EXHIBIT 2

United States Environmental Protection Agency
Office of Water (4607)
Washington, D.C. 20460
EPA 815-R-01-024
August 2001

EPA Statistical Analysis of MTBE Odor Detection Thresholds in Drinking Water

Contents

Tables

Figures

Questions or Comments

Questions or comments about this report may be directed to:

Andrew E. Schulman, Ph.D.
Office of Ground Water and Drinking Water
U.S. Environmental Protection Agency
1200 Pennsylvania Ave. NW, MS 4607
Washington, DC 20460
(202) 260-4197
schulman.andrew@epa.gov

or to the Office of Ground Water and Drinking Water at (202) 260-3022.

Abbreviations

ASTM	American Society for Testing and Materials
MSE	mean squared error
MTBE	methyl tertiary butyl ether
ppb	parts per billion
ppm	parts per million
RFG	reformulated gasoline
SM	Standard Method
SMCL	Secondary Maximum Contaminant Level

Acknowledgments

The following people provided helpful data or discussions about previous MTBE odor studies: Bart Koch, of the Metropolitan Water District of Southern California, for Dale et al. (1998); Jim Prah, of the U.S. Environmental Protection Agency, for Prah et al. (1994); and Yvonne Shen, of the Orange County Water District in Orange County, California, for Shen et al. (1997) and Shen et al. (1998).

The following people acted as external peer reviewers for the report: Gary Burlingame, of the Philadelphia Water Department; Pamela Dalton, of the Monell Chemical Senses Center; Steve Heeringa, of the Institute for Social Research at the University of Michigan; and Pamela Ohman, of the Department of Statistics at the University of Florida. All provided helpful suggestions which led to an improved report.

Executive Summary

As EPA considers establishing a Secondary Maximum Contaminant Level (SMCL) for methyl tertiary butyl ether (MTBE), one consideration is what fraction of people can detect MTBE in drinking water, and how reliably, at given concentrations. At least eight prior studies have attempted to answer this question, but all of these studies have important drawbacks, including small or biased experimental panels, flawed experimental protocols, and erroneous statistical analysis.

This report reexamines the data from one such study, Stocking et al. (2000). Stocking et al. tested 57 subjects for detection of odor from MTBE in bottled water, at concentrations ranging from 2 to 100 parts per billion (ppb). They recruited the largest panel of any MTBE study to date; used a panel of consumers, rather than expert tasters; and used a statistically sound experimental protocol (ASTM method E679-91). For these reasons, the data in Stocking et al. (2000) provide the best available information about MTBE odor detection in drinking water. Stocking et al. (2000) made errors in their statistical analysis, however, which caused them to significantly overestimate some detection thresholds.

Figure ES.1 shows the results of our analysis of the data in Stocking et al. (2000). For any fraction of the population, Figure ES.1 shows the corresponding odor detection threshold, defined as the concentration at which that fraction of subjects can detect MTBE at least half of the time in drinking water. The figure also shows 95% confidence intervals for the thresholds. For example, we estimate that 50% of subjects can detect MTBE at least half of the time at 15 ppb, with a 95% confidence interval of 10 to 22 ppb. This estimate is consistent with those of previous studies, which support an odor detection threshold in the range of 15 to 45 ppb. Table ES.1 compares some of our estimates to those of Stocking et al. (2000). Table ES.2 shows similar results, in terms of the percent of subjects detecting various concentrations at least half the time in drinking water.

In addition to reanalyzing the data from Stocking et al., this report:

- Proposes a general definition of an odor detection threshold, as the concentration at which a certain percent of subjects can detect the contaminant at least a certain percent of the time. We find that 50% is a good choice for the fraction of time, because thresholds defined in this way are easiest to estimate.

- Evaluates two commonly used protocols for odor detection experiments. We find that ASTM protocol E679-91 is statistically sound, but that Standard Method 2150B should not be used because it does not account for the effect of guessing, and allows the probabilities of certain outcomes to depend on the subjects' knowledge of the experimental protocol.

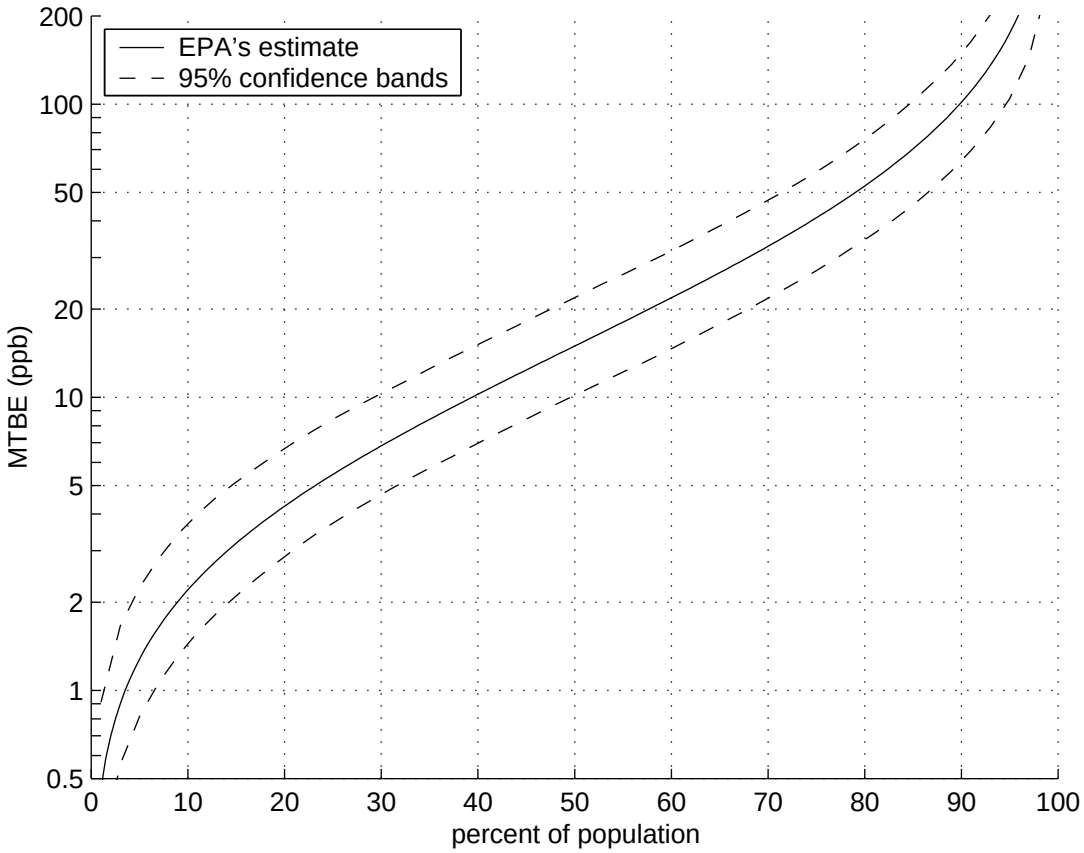

Figure ES.1: MTBE odor threshold estimates: concentrations detectable at least half of the time in drinking water by fractions of the population.

Table ES.1: MTBE odor threshold estimates and 95% confidence intervals (in parentheses), from Stocking et al. (2000) and the present report (EPA).

	Threshold (ppb)	
% of subjects	Stocking et al. (2000)	EPA
5	1.6	1.3 (0.8, 2.3)
10	2.2	2.2 (1.4, 3.7)
25	6.5	5.5 (3.7, 8.5)
50	57	15 (10 , 22)

Table ES.2: EPA estimates of percent of subjects detecting MTBE in at least 50% of samples at various concentrations, with 95% confidence intervals (in parentheses).

MTBE (ppb)	% detecting in at least half of samples
2	9 (4, 14)
5	23 (14, 32)
10	39 (29, 50)
20	58 (47, 68)

- Reviews previous studies of MTBE odor thresholds and identifies problems in each study.
- Evaluates several possible threshold estimators. We find that the simple estimator specified in ASTM method E679-91 performs well when the time fraction being estimated is 50%.

Considerable uncertainty remains in odor threshold estimates, beyond what is reflected in the confidence intervals above:

- Stocking et al. (2000) excluded smokers and subjects over 65 years of age from their experiment. Thresholds estimated from these data may therefore be too low to represent the general population of drinking water consumers. The study also provides insufficient data to allow estimation of the effects of age, disease, or stress, all of which are known to have important effects on odor sensitivity. Larger and more detailed studies will be required to address these questions.
- The thresholds measure only detection of differences between plain and spiked samples, as opposed to recognition of particular odors or rejection of samples as undrinkable. The relationship between these three responses is highly variable and depends in part on a subject's knowledge, beliefs, and tastes.

Nevertheless the data from Stocking et al. (2000) are the best available to date for estimation of MTBE odor detection thresholds.

1. Introduction

Methyl tertiary butyl ether (MTBE) is a fuel oxygenate and octane enhancer for gasoline. Small amounts of MTBE have been added to gasoline in the U.S. since 1979, to replace lead as an octane enhancer. Since 1992, MTBE has been used at higher concentrations in reformulated gasoline (RFG), to fulfill the oxygenate requirements set by Congress in the 1990 Clear Air Act Amendments. RFG is oxygenated gasoline, specially blended to burn cleaner than conventional gasoline, and required to be used year-round in cities with the worst ground-level ozone (smog). In 1997 about 25% of gasoline sold in the United States was reformulated gasoline containing MTBE (US EPA, 2000). At the same time, MTBE has been detected in drinking water wells in areas with leaking underground storage tanks.

As of November 2000, EPA is proposing to establish a secondary maximum contaminant level (SMCL) for MTBE of 5 parts per billion (ppb) in drinking water. The Safe Drinking Water Act (42 U.S.C. §§300f–j) specifies that an SMCL is a non-enforceable standard, intended to maintain the aesthetic quality of drinking water. Section 1401(2) of the Act states that EPA may establish an SMCL for any contaminant that "may adversely affect the odor or appearance of such water and ... cause a substantial number of persons ... to discontinue its use, or ... may otherwise adversely affect the public welfare." In choosing an SMCL, EPA must therefore evaluate the taste and odor properties of MTBE, in particular what fraction of people can reliably detect MTBE in drinking water at concentrations of interest.

Several taste and odor studies of MTBE have been performed. We consider the results of these studies in Section 4. We find that all of the studies performed to date suffer from one or more of the following statistical flaws:

- sample sizes that are too small (4 to 10 subjects) to allow meaningful inferences to the population of drinking water consumers
- subjects selected for high sensitivity to odor
- unclear definitions of odor thresholds; confusion of subject and population thresholds; or failure to treat subject thresholds
- erroneous or unclear statistical analysis
- statistically invalid experimental protocols

Confusion about the exact meaning of an odor threshold is especially problematic and widespread.

In addition to the "within-study" problems listed above, there are "across-study" inconsistencies which make interpretation harder. Studies test different ranges of concentrations; use

different experimental protocols, with additional minor variations; and use different statistical methods to analyze their results.

This report has two goals. First, we aim to resolve as many as possible of the problems described above, by clarifying the meaning of odor detection thresholds, and by identifying an experimental protocol and methods of statistical analysis which allow the most reliable statistical inferences for odor thresholds. We believe that the methods which we identify here would make a good standard approach for defining and estimating odor thresholds. Second, we apply these techniques to the best available MTBE odor data set, that of Stocking et al. (2000), in order to estimate odor detection thresholds for MTBE in drinking water.

Studies of taste or odor must distinguish between detection, recognition, and rejection of a contaminated sample. A consumer may detect that one sample is different from another, but not recognize an odor or reject either sample as undrinkable. In this report we consider only detection thresholds, for two reasons. First, existing studies of MTBE consider mostly detection, with a small amount of data on recognition and none on rejection. Second, whether rejection follows detection depends on a subject's knowledge, beliefs, and tastes, all of which are outside our scope.

We begin in Section 2 by offering a precise definition of odor thresholds, based on a general model of odor detection. In Sections 3 and 4, we review standard odor detection protocols and existing studies of MTBE odor thresholds, in light of our definition. In Section 5 we reanalyze the data from Stocking et al. (2000), to estimate MTBE odor thresholds. In Section 6 we compare the performance of several threshold estimators under the conditions of Stocking et al.'s experiment. In Section 7 we present our conclusions.

2. A General Model of Odor Detection

In this report we suppose that each person has a probability of detecting a substance of interest at each concentration, and that the probability of detection increases with concentration. "Detection" means the observation of a difference in taste or odor between a sample of plain water (of whatever type; see below) and a sample of water plus the substance of interest. For convenience we refer to the substance of interest as a "contaminant."

Figure 2.1 shows three hypothetical dose-response curves that would satisfy our model. Given a concentration, each curve determines the subject's probability of detecting the contaminant. For example, at 10 ppb, each of the hypothetical curves assigns probability $1/2$ of detection. One of the curves (sold line) yields zero probability of detection at concentrations below about 2 ppb, and probability one of detection above about 50 ppb. The other two curves never assign probabilities of exactly zero or one; the probabilities only approach zero or one for very small or large concentrations (although this is not always clear from the graph). This property is typical of statistical dose-response models.

Our model implies that each subject has a set or range of odor thresholds, determined by the probabilities of detection. For example, under any of the dose-response curves in Figure 2.1, the subject has a 50% detection threshold of 10 ppb. The 75% detection threshold is somewhere between about 11 and 25 ppb, depending on the dose-response model. In general, for a number T between 0 and 1, a subject's $(100T)$% *detection threshold*, denoted C_T, is the concentration at which the subject can detect the contaminant $(100T)$% of the time, or in $(100T)$% of samples. We call T the *time fraction* of the odor threshold. A dose-response curve is therefore a function that determines the time fraction of detections as a function of dose. As observed above, under most statistical dose-response models there is no 0% or 100% threshold,

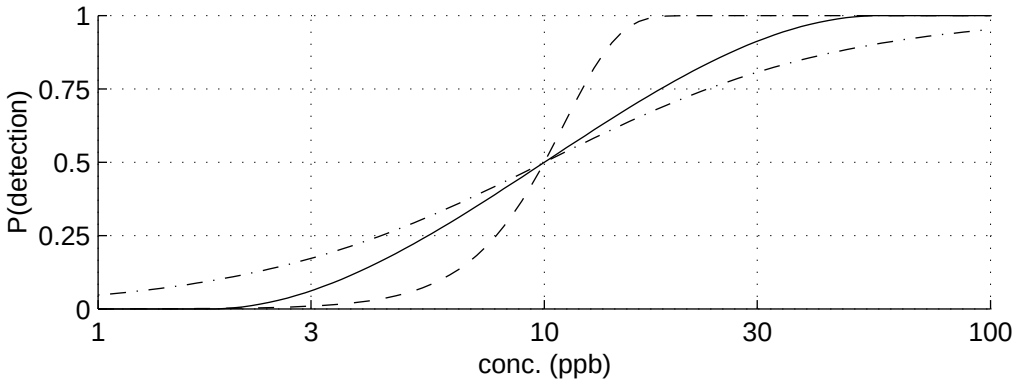

Figure 2.1: Possible dose-response curves for a hypothetical subject.

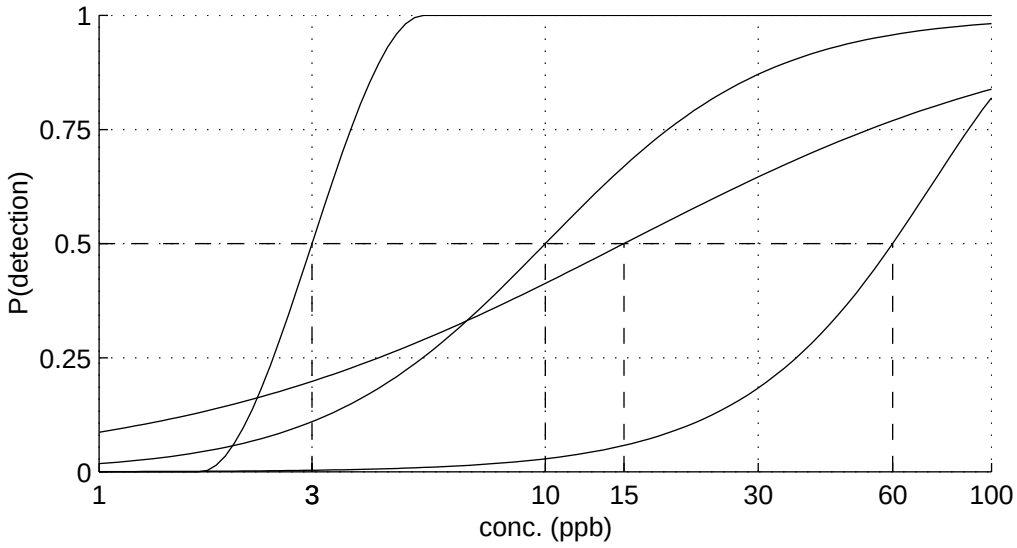

Figure 2.2: Dose-response curves and corresponding 50% detection thresholds for a hypothetical population of four subjects.

only 99%, 99.9%, 99.99%, ... and 1%, 0.1%, 0.01%, ... thresholds.

Consider now a population of subjects, each with his or her own odor dose-response curve. Figure 2.2 shows the dose-responses curves and corresponding 50% detection thresholds for a small hypothetical population. The 50% thresholds in Figure 2.2 range from 3 to 60 ppb. In general, given a population of subjects and numbers S and T between 0 and 1, we define the population's (S, T) *odor detection threshold*, denoted $C_{S,T}$, as the $(100S)$-th percentile of the individual subjects' $(100T)$% odor detection thresholds. That is, $(100S)$% of the subjects can detect the contaminant $(100T)$% of the time or in $(100T)$% of samples, at or below $C_{S,T}$. For example in Figure 2.2, the 50% subject thresholds are 3, 10, 15, and 60, so the $(50\%, 50\%)$ population threshold is anywhere between 10 and 15. We call S the *subject fraction*, and T as before is the time fraction. $C_{S,T}$ may be estimated as a sample $(100S)$-th percentile of $(100T)$-percentage points of estimated dose-response curves.

Other definitions of thresholds are of course possible. Different statistics may be used to summarize over the time or subject fractions; for example, most studies summarize over subjects by computing a geometric mean of individual thresholds. One can also summarize first over subjects, then over time. Dale et al. (1998) take this approach by first computing the dose response for an "average" subject, that is, by averaging dose-response curves over subjects, and then choosing probability (time fraction) thresholds of the averaged response curve.

We believe that our definition of odor thresholds is the most useful one, because it goes directly to the true objective of the analysis: what percent of people will detect the contaminant some given fraction of the time. To be useful, other definitions have to be interpreted in terms of these fractions. For example, an average subject threshold does not tell us directly how many

people can detect the contaminant at that level; to get that information we have to interpret the average as approximately a median. This is even more true for a geometric mean, which is a less familiar statistic, and sometimes has to be explained as approximately a median.

Regardless of the particular definition, however, a crucial point is that any population odor threshold statistic summarizes the dose-response curves over *both the subject and time fractions*. That is, a threshold must correspond at least implicitly to some percent of people detecting some percent of the time. In order to interpret the threshold, we have to know what the fractions are. Although this point may seem obvious, many of the studies which we review in Section 4 are difficult to interpret precisely because they fail to state or evaluate how they summarized the dose-response, in particular over time. A typical claim is that at some concentration, some percent of subjects "can reliably detect" the contaminant. In some cases the time fraction is implicit in the experimental protocol or threshold estimator, but in other cases it cannot be determined from the information in the study.

Our odor dose-response model does not account for several factors which are known to influence odor detection. Covariates such as age, disease, smoking status, and stress have all been shown to have important effects on odor sensitivity (Schiffman, 1992; Smith and Duncan, 1992). We do not treat these covariates here because, with the exception of smoking status, no available studies of MTBE provide information about them. In the case of smoking status, all MTBE studies which mention it exclude smokers; this approach obviates any modeling of smoking effects, at the cost of introducing a negative bias in the threshold estimates.

Our model also does not consider desensitization to odor through time. In serial sniffing tests, odor sensitivity is likely to decrease for a period of minutes to hours following each sniff. Since we are interested in drinking water, the ideal rest period between test sniffs would be the mean time between a consumer's drinks of tap water, or on the order of one hour. In practice, cost constraints require that tests progress faster, on the order of one minute between sniffs. This faster testing probably reduces odor sensitivity, but most studies provide no data about this question, so we do not consider it here.

Finally the odor threshold is affected by the medium, that is, the type of water in which the contaminant is presented. A contaminant may be more easily detected in distilled water, for example, than in tap water, which typically contains chlorine and other contaminants that may mask the odor of the contaminant of interest. For this reason any dose-response curve must be assumed to be valid only for the stated medium. Fortunately all studies of MTBE clearly state which medium is used, and some (Shen et al., 1997) test more than one medium.

3. Odor Detection Protocols

Three experimental protocols, Standard Method 2150B (APHA, 1995) and ASTM methods E1432-91 and E679-91 (ASTM, 1995a,b), specify standard methods for performing odor detection experiments and estimating odor detection thresholds. In this section we describe the experimental design and threshold estimators of each protocol. We argue that Standard Method 2150B, while it has some advantages, has statistical flaws that make it useless for statistical modeling and estimation. The ASTM protocols avoid these problems and should therefore be preferred for statistical modeling.

Samples of water plus the contaminant (substance of interest) are referred to below as "contaminated" or "spiked." Again the language of contamination is used for convenience and without prejudice. Samples of plain water are referred to as "uncontaminated" or "blanks."

3.1 ASTM Methods E679-91 and E1432-91

The ASTM protocols specify that samples of the contaminant are to be presented to subjects in increasing order, using a "forced choice triangle test." That is, for each concentration, the subject is presented with three bottles, two blank and one spiked (or in some versions, one blank and two spiked). The subject is asked to sniff each of the three bottles and identify one as different from the other two. One bottle must be selected each time, so if the subject cannot detect a difference between the bottles, s/he is forced to guess. Thus the subject has one chance in three of choosing the correct bottle, even if s/he cannot detect a difference. When analyzing the results of a forced-choice test, one has to distinguish between the probability T of detecting a difference, and the probability P of answering correctly. P is greater than T, because a subject who cannot detect the contaminant may still guess correctly. The relationship between the two may be shown to be

$$P = T + (1 - T)C \tag{3.1}$$

where C is the probability of a correct guess. For a forced-choice triangle test, $C = 1/3$ and so $P = (1/3) + (2/3)T$.

The two ASTM protocols specify different subject threshold estimators. ASTM method E1432-91 defines an individual odor threshold as "the concentration for which the probability of detection of the stimulus is 0.5" (ASTM, 1995a). Method E1432-91 proposes to estimate this threshold by using nonlinear regression to estimate the dose-response relationship of "percent correct above chance" (i.e., corrected for the 1/3 probability of a correct guess) to log-dose, then finding the dose which yields 50% response. This method is said to be valid only for an experiment with multiple presentations of each concentration and a total of 20 or more

presentations per subject. A population threshold is then estimated as the geometric mean of individual thresholds, or, if desired, a $(100S)\%$ normal-theory quantile estimate of individual thresholds. In the latter case, ASTM method E1432-91 seeks to estimate the $(S, 0.5)$ population threshold.

ASTM method E679-91 uses a simpler individual threshold estimator, intended for smaller experiments where the dose-response relationship is harder to estimate. Using this method, each subject's threshold is estimated as the geometric mean of the highest concentration at which the subject gave a wrong answer and the next higher tested concentration. (If the subject gave all right or all wrong answers, s/he is assumed to have a given a wrong answer at a next lower concentration in the sequence, or a right answer at a next higher concentration.) As an example, the answers at different concentrations of MTBE from subject #40 in Stocking et al. (2000) looked like this:

conc. (ppb)	2	3.5	6	11	19	33	57	100
answer	o	o	+	o	+	+	+	+

where o represents a wrong answer, and + a right answer. Since this subject's last wrong answer occurred at 11 ppb, his or her odor threshold is estimated as the geometric mean of 11 and 19, or 14. Population thresholds are then estimated as the geometric mean of the subject thresholds. Note that this method does not explicitly take into account the effect of guessing, and it only uses data at the highest concentrations for each subject. Even so, we will see in Section 6 that it performs fairly well at estimating the 50% subject threshold.

3.2 Standard Method 2150B

Standard Method 2150B defines a subject's odor threshold as the concentration at which odor is "just detectable" (APHA, 1995). The method states that because of variability there is no absolute odor threshold, but it does not name a probability of detection that corresponds to just-detectability.

Standard Method 2150B specifies that eight contaminated samples are to be presented one at a time to each subject, in ascending order of concentration, with two or more blanks mixed in at random. Subjects are asked to state whether they detect any odor in each sample. A detection threshold is then estimated for each subject as the lowest concentration at or above which the subject made no mistakes in identifying samples with or without contaminant. (This definition is not entirely clear however, because an accompanying graphic shows an arrow pointing to the *midpoint* between the last miss and the next tested concentration.) Group thresholds are estimated by using "appropriate statistical methods" to compute the "most probable average threshold" from the individual thresholds; geometric means are recommended for this purpose.

Standard Method 2150B has some advantages over the ASTM methods. One advantage is that it requires fewer sniffs: only 10 sniffs for 8 contaminated samples, compared to 24 sniffs using the triangle test. This may lead to more accurate results by reducing desensitization of the nose over time. Another advantage is that when a subject cannot detect the contaminant, it

is more natural for him or her just to say so than to be forced to guess, as in the forced-choice test. It is possible that a subject who is annoyed or distracted by guessing would give less reliable answers.

On the other hand, Standard Method 2150B presents statistical difficulties which the ASTM methods do not. The problem is that under this method, the probabilities of the various outcomes may depend on the subject's knowledge of the experimental protocol. To see this, consider first the ASTM protocols, which use the forced-choice triangle test. Under this test, a subject who knows the details of the experimental protocol gains no information to help him or her take the test. In each presentation, two bottles are plain water, and one is spiked; the subject may know this—in fact he or she should know it—but gains no information thereby about which bottle he or she should choose. Moreover, when a subject cannot detect the contaminant, the probability of a correct answer is known: it's $1/3$. By contrast, under Standard Method 2150B, a subject who knows the experimental protocol gains important information about the test: he or she learns that of, say, 10 bottles presented, 8 are contaminated and 2 are not. Thus when asked whether he or she detects odor in the sample, the subject knows that, 4 times out of 5, the "right" answer is yes. This poses several problems for the modeler. First, when a subject cannot detect the contaminant in a spiked sample, is the probability of correctly answering "yes" anyway by guessing really 80%? If so, the power of discrimination of the test will be poor; and if not, then what is the probability? Second, when presented with a blank, what is the probability that a subject will identify it as contaminated? Is a subject who knows the protocol more likely to imagine smelling contaminant in a blank? If so, how much more likely? And what percent of subjects knew the protocol? (For reasons of economy, subjects in odor threshold studies are often lab employees or even study authors, who take such tests repeatedly and know the details of the experimental protocol.)

The answers to these questions are unknown, in the absence of a separate experiment designed to measure them. Standard Method 2150B avoids answering them by specifying a simple threshold estimator which does not require that one know the answers. Yet no statistical justification is given for preferring this estimator, and indeed no such justification is possible without answering the above questions, because the statistical properties of the estimator depend on unknowns. Nor is it possible to propose other, model-based estimators, without being able to assign probabilities to right or wrong guesses. For these reasons, we believe that Standard Method 2150B should not be used for odor detection experiments.

4. Review of Previous Studies of MTBE

In this section we review previous studies of MTBE odor detection thresholds in drinking water. Although each of the studies has some methodological problems, taken together they present a fairly consistent picture of MTBE odor thresholds, which we discuss at the end of the section.

4.1 TRC (1993) and API (1994)

TRC (1993), in a report to ARCO Chemical Company, estimated odor detection and recognition thresholds for MTBE in distilled water. They tested five concentrations ranging from 106 to 1,483 ppb, with two replicates on a panel of seven subjects. Although the authors claimed to use Standard Method 2150B, they used a forced-choice triangle test, and to estimate the individual detection thresholds they combined the results of that test with statements from the subjects about whether they could detect a difference.

API (1994) also estimated odor thresholds for MTBE in distilled water. The authors tested six concentrations ranging from 23 to 740 ppb, with two replicates on a panel of seven subjects. This study was also carried out by TRC, and is quite similar to TRC (1993). The methodology of the two studies is the same, and three of the panel members were apparently the same.

API (1994) defines the population detection threshold as the smallest concentration at which 50% of the population can detect the contaminant, but does not say how reliable the detection should be. That is, the authors specify a population fraction of $S = 0.50$, but fail to identify the time fraction T. A complicated method is used to estimate the population threshold. A linear regression model is fit with response equal to the midpoints between the tested log-concentrations, and predictor equal to the rank of the number of times the contaminant was first detected at each concentration, standardized and transformed to a normal score. The intercept of the regression is claimed to be the log-detection threshold.

Using this method, TRC (1993) and API (1994) find group odor detection thresholds of 95 and 45 ppb, respectively, for MTBE in distilled water.

These two studies present at least three problems. First, the sample sizes are small: each study used only seven subjects, and the two studies combined used a total of 11 subjects. Second, the higher threshold estimate in TRC (1993) is almost certainly due to the higher range of concentrations tested there. The two studies shared the same methodology and even half of their experimental panels; the only apparent difference between them was in the ranges tested. Moreover the estimated threshold in TRC (1993) lies below the range of observation, implying that an extrapolation was used and casting doubt on the result. Third, the statistical model used for the regression is theoretically unsatisfactory, because the random error occurs

in the predictors (sample percent correct) rather than the responses (log-concentrations). The problem is aggravated by correlation among the predictors, which arises because they are all estimated from the same set of subjects. As a result the regression parameter estimates are less reliable than they could be.

4.2 Prah et al. (1994)

Prah et al. (1994) tested 37 subjects for detection of MTBE in distilled water, at concentrations ranging from 31 ppm (parts per million, by volume) to 1000 ppm. They used an experimental protocol similar to that of Standard Method 2150B, in which 6 contaminated samples and 2 blanks were presented to subjects one at a time. They found an odor threshold of 180 ppb, but this number represents the concentration of MTBE in the head space (air) above the water in the sample jar, not the concentration in water (J. Prah, personal communication). The corresponding concentration in water was presumably 31 ppm, which, while the lowest concentration tested, is 3 to 4 orders of magnitude greater than the concentrations at which subjects have detected MTBE in other studies. Since 31 ppm was the lowest concentration tested in this study, the results cannot plausibly shed light on detection thresholds in the range of 2 to 100 ppb.

4.3 Young et al. (1996)

Young et al. (1996) estimated taste and odor thresholds of 59 contaminants, including MTBE, in bottled mineral water. For MTBE they used 9 female subjects between 25 and 55 years of age, selected for "above average" odor sensitivity. Young et al. (1996) used a modified version of the HMSO odor-detection protocol (HMSO, 1982). For each subject they first identified a provisional odor threshold, by presenting spiked samples in increasing concentrations, each paired with a blank, until the subject picked out the right sample. They then verified the provisional threshold by a sequence of further trials with blank-blank and spiked-blank pairs, at the provisional and next higher thresholds. The procedure seems designed to pick out with some certainty the lowest concentration at which a subject can detect the contaminant, but the authors do not identify a detection probability that they associate with just-detectability, nor do they attempt to determine the effective time fraction of their estimator.

Young et al. (1996) find a geometric mean MTBE odor threshold of 34 ppb. Unfortunately, the range of tested concentrations in Young et al. cannot be determined from the paper; although the authors list the dilutions they used of their stock MTBE solution, they do not state the concentration of MTBE in the stock solution.

4.4 Shen et al. (1997) and Shen et al. (1998)

Shen et al. (1997) estimated odor thresholds of MTBE in odor-free, tap, and chlorinated odor-free water. They used Standard Method 2150B, with 4 replicates for odor-free water and 2

each for tap and chlorinated water. Panels for each replicate consisted of 8 to 10 "experienced" subjects; many of the panelists took part in two or more replicates. Concentrations of 2.5 to 150 ppb were tested. At room temperature, individual detection threshold ranges were 2.5–100 in odor-free water, 2.5–150 in tap water, and 5–150 in chlorinated water. Geometric means of subject thresholds ranged from 13.5 to 40.3 for odor-free water, 13.5 to 33.9 for tap water, and 31.3 to 43.5 for chlorinated water. Additional tests at higher temperatures (Shen et al., 1997) and at various temperatures and chlorine concentrations (Shen et al., 1998) gave similar results.

As we argue in Section 3.2, because of the design of Standard Method 2150B, the meaning of Shen et al.'s results is unclear. Moreover both studies modified the Standard Method by discarding all responses from a subject in a trial if the responses resulted in an "anomaly," that is, identifying a blank as contaminated, or failing to detect odor in the highest concentration of contaminant after identifying a lower concentration. While this practice may be defensible as a way of improving the estimate by removing subjects who are only guessing, it illustrates the problem of using a method such as SM 2150B, which does not account for guessing. A well-designed experimental protocol should not require that one throw out otherwise validly obtained data.

4.5 Dale et al. (1998)

Dale et al. (1998) estimated odor thresholds of MTBE in odor-free water. They used a panel of 4 to 7 trained sniffers, all of whom were required to be able to detect "off-flavors and off-odors at very low concentrations." Subjects took triangle tests at 7 concentrations ranging from 6 to 118 ppb, with 6 replicates at each concentration for each subject.

Dale et al. (1998) used probit regression with a common slope (B. Koch, personal communication; see Section 6.2 for a description) to estimate dose-response curves for each subject. They then summarized the curves as odor thresholds in two ways. First, they averaged the dose-response curves over subjects, to find the dose-response for an "average" subject. Based on this function, the average subject could detect MTBE at 20 ppb about 50% of the time, for example. In a second, separate analysis, they found that 60% of subjects "would perceive the MTBE" odor at 43–71 ppb (95% confidence interval), but did not say how often. That is, they specified the subject fraction $S = 0.60$, but failed to state the time fraction T.

Dale et al. (1998) used a small panel, intentionally biased toward more sensitive subjects. It is difficult to draw inferences from such a panel to the population of drinking water consumers. Their estimated dose-response functions also appear to be wrong, because they modeled the probability of odor detection as a function of untransformed concentration, rather than log-concentration. As a result, the averaged dose-response curve shows approximately a 45% probability of detection at a concentration of zero. The authors also did not account for the effect of guessing. Finally their threshold estimates are hard to interpret, because they either describe a hypothetical average subject, or do not specify the time fraction of the threshold.

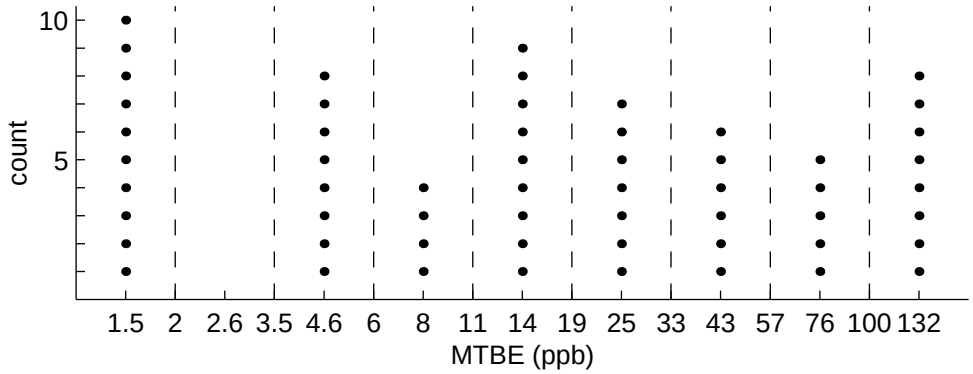

Figure 4.1: ASTM odor threshold estimates for 57 subjects (Stocking et al., 2000). Dots are thresholds; dashed lines are the tested concentrations.

4.6 Stocking et al. (2000)

Stocking et al. (2000) estimated odor thresholds of MTBE in bottled water at room temperature. They used a consumer panel of 57 subjects, the largest of any MTBE study to date. The experiment followed ASTM protocol E679-91, including the forced-choice triangle test, with 8 concentrations, equally spaced on the log scale, from 2 to 100 ppb. Each concentration was presented once to each subject, although subjects were allowed to repeat each presentation once if they were unsure of their answer the first time. Subjects were not selected from the entire population of drinking water consumers, but rather were drawn from a list of 10,000 consumers maintained by the National Food Laboratories for taste and odor experiments. The panel was balanced by sex and age within the range of 18 to 65 years of age. Smokers were excluded. Further details about subject selection and training, container preparations, test administration, and quality control are provided in Malcolm Pirnie (1998). The data from this experiment are presented in Appendix A.

Individual odor thresholds were estimated according to ASTM protocol E679-91. Figure 4.1 shows the set of 57 threshold estimates; in accordance with the ASTM protocol, the estimates fall between the tested concentrations. For example, subjects who failed to identify MTBE at 11 ppb, but correctly identified it at 19 ppb and all higher concentrations, were assigned thresholds equal to the geometric mean of 11 and 19, or 14. Ten subjects correctly identified MTBE at all tested concentrations, from 2 to 100 ppb; these subjects were assigned detection thresholds of 1.5 ppb. Eight subjects failed to identify MTBE at 100 ppb (although all identified some lower concentrations correctly), and were assigned thresholds of 132 ppb. The geometric mean of the individual threshold estimates is 15 ppb.

In a separate analysis, Stocking et al. (2000) used a logistic regression model

$$\log\left(\frac{P}{1-P}\right) = a + b\log C \tag{4.1}$$

to model the fraction P of the population correctly identifying MTBE at concentration C (in

ppb). The data used to fit this model are shown on the last line of Appendix A. For example, at 2 ppb the sample fraction is 25/57. The slope and intercept parameters were then estimated as $a = -0.726$ and $b = 0.569$. (Stocking et al. did not list their parameter estimates; EPA used maximum likelihood (McCullagh and Nelder, 1989) to derive the estimates given here. These estimates give probabilities that agree fairly well with Tables 8 and 9 of Stocking et al.) The authors then corrected for the effect of guessing in the forced-choice test by adding 1/3 to the probability of detection, to approximate the probability of a correct answer. For example, for 25% probability of detection, take $P \simeq 1/3 + 0.25 = 0.583$ and solve (4.1) for C to find $C = 6.5$ ppb (Stocking et al. find $C = 6.2$). The authors' interpretation is that 25% of subjects can detect MTBE at about 6 ppb.

The logistic regression analysis has several problems. First, the correction for guessing contains a simple error in the approximation of $P = 1/3 + (2/3)T$ (equation (3.1)) by $P = 1/3+T$. This approximation is accurate enough when T is small, but it makes a large difference when $T = 50\%$. For example using $P \approx 1/3 + T$, Stocking et al. find a threshold when $T = 50\%$ of about 57 ppb; using $P = 1/3 + (2/3)T$, the threshold is 12 ppb.

A second problem lies in Stocking et al.'s interpretation of the regression results. The authors use the results of the fitted model to conclude that "2 μg/L, 6 μg/L, and 57 μg/L ... represents the concentration [sic] at which 5%, 25%, and 50%, respectively, of the subjects can make accurate discriminations." Note that no time fraction is stated to correspond to "accurate discriminations." In fact, the stated percentages are a combination of subject and time fractions, in such a way that neither fraction can be inferred from them. To see this, consider the following two possible descriptions of the study population:

1. At 12 ppb, 20% of the subjects have a 10% chance of detecting MTBE, and the other 80% have a 60% chance of detecting it. On average, the fraction of detections at 12 ppb will be $(0.20)(0.10) + (0.80)(0.60) = 0.50$, or 50%.

2. At 12 ppb, 20% of the subjects have a 90% chance of detecting MTBE, and the other 80% have a 40% chance of detecting it. On average, the fraction of detections at 12 ppb will be $(0.20)(0.90) + (0.80)(0.40) = 0.50$, or 50%.

Both of these descriptions agree with Stocking et al.'s (corrected) regression results of 50% detections at 12 ppb, but conflict with their interpretation. In the first case, 80% of subjects are at least 50% likely to detect at 12 ppb; while in the second case, only 20% are. Stated another way, in the first case 12 ppb is an (80%, 50%) threshold, while in the second case it is a (20%, 50%) threshold. Although this example is artificial, it shows that Stocking et al.'s interpretation in terms of subject fractions is not supported by their analysis.

Another problem with model (4.1) is that the data in Appendix A do not represent 57×8 independent observations, since repeated observations on a single subject are correlated. Stocking et al. (2000) apparently did not account for this correlation in estimating a and b. The result may be less efficient estimates and confidence intervals that are too narrow. This may be a reasonable compromise against the difficulty of fitting a model, but the choice needs to be evaluated. In Section 6 we consider some models that include intra-subject correlation.

Finally, fitting a model such as (4.1), which does not account for guessing, then adjusting for guessing post-hoc as in Stocking et al. (2000), simplifies the analysis but may not be the best approach. A better method might be to fit a model which explicitly takes the probability of guessing into account. We consider this possibility in Section 6.

4.7 Summary

Table 4.1 summarizes the results of existing odor studies of MTBE, as well as the problems that we have identified above. All of the studies have important drawbacks; small panels and unclear threshold definitions or properties are the most common.

Despite the problems that we have detailed here, the results listed in Table 4.1 are fairly consistent and support a $(50\%, 50\%)$ odor detection threshold for MTBE somewhere in the range of 15 to 45 ppb. We do not include the results of TRC (1993) and Prah et al. (1994) in this range, since, as we argued above, the higher threshold estimates of these studies appear to have been caused by the higher concentrations that they tested. Our reanalysis of the data from Stocking et al. (2000), in Section 5, finds a $(50\%, 50\%)$ odor detection threshold of 15 ppb, consistent with these findings.

Table 4.1: Summary of MTBE odor threshold studies.

	TRC (1993)	API (1994)	Prah et al. (1994)	Young et al. (1996)	Shen et al. (1997)	Dale et al. (1998)	Stocking et al. (2000)
type of water	distilled	distilled	distilled	mineral	tap[a]	odor-free	bottled
number of subjects	7	7	37	9	8–10	4–7	57
range tested (ppb)	106–1,480	23–740	3.1×10^4 -1×10^6	?	2.5–150	6–118	2–100
threshold estimate (ppb)	95	45	180[b]	34	14–34	20	15
tested range too high	*		*				
small panel ($\leqslant$ 10)	*	*		*	*	*	
expert panel				*		*	
flawed experimental protocol			*		*		
unknown threshold properties	*	*	*	*	*	*	*
errors in analysis	*	*				*	*

[a] results were similar in distilled water, and somewhat higher in chlorinated water
[b] in air above sample

5. EPA's Analysis of Stocking et al. (2000)

In this section we reanalyze the data from Stocking et al. (2000), in order to estimate MTBE odor detection thresholds in drinking water. Stocking et al. (2000) contains the best available data for estimating MTBE odor thresholds: it used the largest panel of any study to date; used a panel of consumers rather than expert sniffers; followed the ASTM protocol, in particular the triangle test; and tested a low enough range of concentrations, centered approximately around the threshold estimates of other MTBE studies. The data from the study are also reproduced in full in the paper; we reproduce them here in Appendix A.

The data from Stocking et al. (2000) do have some disadvantages:

- The study panel was not selected as a random sample of drinking water consumers, as we assume below. Rather, subjects were chosen from a list, maintained by the National Food Laboratories, which conducted the experiment, of self-selected taste-and-odor study participants. It is therefore difficult to know how far any analysis of this data set can be generalized to the population of drinking water consumers. We know of no reason to expect any systematic bias due to this selection process, however.

- Only non-smokers ages 18 to 65 were included in the study. This probably has the effect of biasing the results toward low thresholds. The panel was also balanced, apparently deliberately, by gender and by three age groups. It is not known to what extent these balances reflect the population of drinking water consumers.

- Only one replicate of each concentration was tested on each subject. Individual odor thresholds can therefore be difficult to estimate, as we see in Section 6.1. This variability is accounted for in our confidence intervals, however.

While keeping these reservations in mind, we believe that the size and careful design of the experiment make these data suitable for estimation of MTBE odor thresholds.

Our goal is to estimate $(S, 50\%)$ odor detection thresholds for MTBE, with S ranging from 0 to 100%. That is, we want to know at what concentration any given fraction of the population can detect MTBE at least half of the time. Conversely, given a concentration of MTBE, we want to know what fraction of people can detect it at least half of the time. Although we could choose any time fraction to estimate, 50% seems a reasonable choice. The ASTM estimator also tries to estimate 50% thresholds, and the results of Section 6 show that 50% thresholds are easiest to estimate.

In Section 6 we evaluate several estimators of population thresholds, by simulating their effectiveness under conditions similar to those of Stocking et al. (2000). When the time fraction of the estimand is 50%, a simple estimator with small bias and variance may be obtained by

Table 5.1: Population threshold estimates and 95% confidence intervals (in parentheses), from Stocking et al. (2000) and the present report.

		Threshold (ppb)	
% of subjects	% of samples	Stocking et al. (2000)	EPA
5	50	1.6	1.3 (0.8, 2.3)
10	50	2.2	2.2 (1.4, 3.7)
25	50	6.5	5.5 (3.7, 8.5)
50	50	57	15 (10 , 22)

computing an ASTM threshold estimate for each subject (following ASTM (1995b); see Section 3.1 or 6.1), then computing a lognormal quantile estimate from the set of subject thresholds. Let $\hat{C}_1, \ldots, \hat{C}_{57}$ be the subject threshold estimates; these are tabulated in Appendix A. Let $\hat{\mu}$ and $\hat{\sigma}$ be the sample mean and standard deviation, respectively, of $\log \hat{C}_1, \ldots, \log \hat{C}_{57}$; we find $\hat{\mu} = 2.71$ and $\hat{\sigma} = 1.49$. Then for any given S, the threshold estimate is

$$\hat{C}^{\text{ASTM}}_{S,0.5} = \exp\bigl(\hat{\mu} + \hat{\sigma}\,\Phi^{-1}(S)\bigr) \tag{5.1}$$

where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution.

Figure 5.1 shows a graph of $\hat{C}^{\text{ASTM}}_{S,0.5}$ as a function of S, with pointwise 95% confidence bands, obtained by bootstrap (Efron and Tibshirani, 1993). The subject threshold estimates are also plotted for comparison. Table 5.1 lists the population threshold estimates at several values of S, as well as the corresponding confidence intervals from Figure 5.1, and corresponding estimates from Stocking et al. (2000).

In Table 5.1, the large difference between EPA's and Stocking et al.'s 50% threshold estimates is mainly due to Stocking et al.'s approximation error, described in Section 4.6. At lower subject fractions the magnitude of that error is small, and Stocking et al.'s analysis, although we argued that it was not the right one, gives similar results to ours.

In order to estimate the fraction of people detecting a given concentration C at least 50% of the time, one can invert (5.1) to get

$$\hat{S}_C = \Phi\bigl((\log C - \hat{\mu})/\hat{\sigma}\bigr). \tag{5.2}$$

Some results are shown in Table 5.2, with bootstrap confidence intervals. The plot of $\hat{S}_C$ as a function of C is the same as Figure 5.1, but going from the vertical to the horizontal axis.

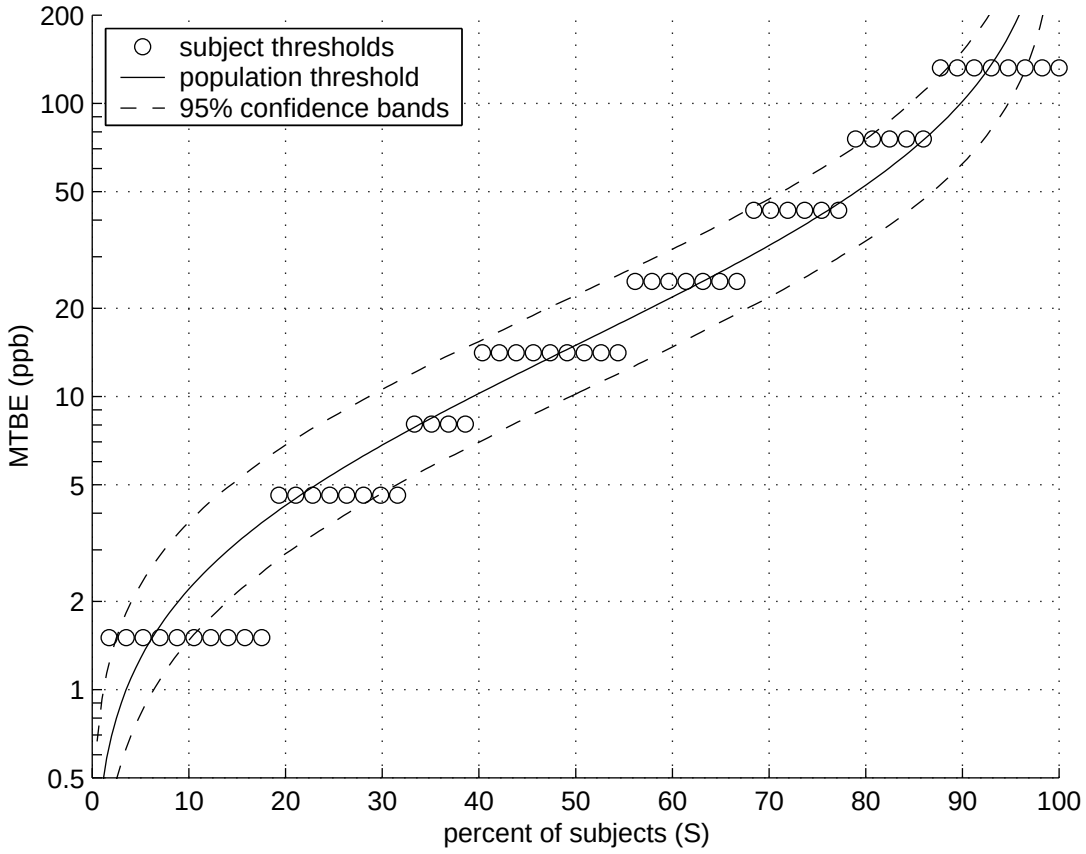

Figure 5.1: EPA estimates of population thresholds $C_{S,0.5}$, with 95% pointwise confidence bands. Circles are ASTM subject threshold estimates, plotted against their ranks expressed as a percentage of 57.

Table 5.2: EPA estimates of percent of subjects detecting various concentrations at least 50% of the time, and 95% confidence intervals (in parentheses).

MTBE (ppb)	% detecting
2	9 (4, 14)
5	23 (14, 32)
10	39 (29, 50)
20	58 (47, 68)

6. Comparison of Odor Threshold Estimators

In this section we compare the performance of several odor threshold estimators. We begin in Section 6.1 with single-subject thresholds, then consider population thresholds in Section 6.2.

The performance of threshold estimators in an odor experiment depends on, among other things, the experimental protocol, panel size, number of concentrations tested, and number of replicates. In order to shed the most light on our analysis of the data in Stocking et al. (2000), we restrict our comparison to the conditions of that study: a forced-choice triangle test of eight contaminant concentrations, with 57 subjects and one replicate per subject per concentration.

6.1 Subject Thresholds

Consider first an odor detection experiment on a single subject. Eight sample concentrations $c_1, \ldots, c_8$, equally-spaced on the log scale, are presented once each, in increasing order, in a forced-choice triangle test. For any given concentration c, let

$$\begin{aligned} t &= \text{the probability that the subject detects the contaminant} \\ p &= \text{the probability that the subject chooses the correct bottle} \\ y &= \begin{cases} 1 & \text{if the subject chooses the correct bottle,} \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{6.1}$$

We assume that if the subject detects the contaminant (probability t), s/he chooses the correct bottle; but if s/he fails to detect the contaminant (probability $1-t$), s/he still has probability $1/3$ of choosing correctly anyway by guessing. Therefore, the relationship between t and p is

$$p = t + (1-t)(1/3) = (1/3) + (2/3)t. \tag{6.2}$$

The expected value of y (given c) is p.

Consistent with our general model of odor detection in Section 2, we also assume that $t = g(c)$, for some increasing function g that maps any concentration c onto a probability t.

The purpose of the experiment is to estimate the subject's $(100T)\%$ odor threshold, for some number T between 0 and 1. That is, we want the concentration C_T at which the subject can detect the contaminant $(100T)\%$ of the time or in $(100T)\%$ of samples. C_T is therefore the solution to $T = g(C_T)$.

6.1.1 Estimators

We have already defined one estimator of C_T in Section 3.1: the ASTM estimator, which we denote $\hat{C}_T^{\text{ASTM}}$, is the geometric mean of the highest concentration at which the subject failed to identify the contaminant, and the next higher tested concentration. For the purposes of this estimator we extend the sequence of concentrations $c_1, \ldots, c_8$ to $c_0, \ldots, c_9$, so that the new sequence is still evenly spaced on the log scale, and we assume that the subject would have answered incorrectly at c_0 and correctly at c_9. So for example, if a subject answers correctly at all of $c_1, \ldots, c_8$, then his or her ASTM threshold is the geometric mean of c_0 and c_1.

ASTM protocol E679-91 states that $\hat{C}_T^{\text{ASTM}}$ is an estimator of a subject's 50% detection threshold. That is, it is valid only when $T = 0.5$. There is no provision for modifying the estimator for other values of T.

Other estimators of C_T may be obtained by fitting a dose-response model

$$\begin{aligned} y_j &\sim \text{Bernoulli}(p_j) \quad \text{indep.} \\ p_j &= t_j + (1 - t_j)K \\ t_j &= h(a + b \log c_j) \end{aligned} \tag{6.3}$$

to the data $(c_1, y_1), \ldots, (c_8, y_8)$, and reading concentrations from the fitted model. K is a number that describes the probability of guessing, and h is a function, called the *link function*, which maps any real number onto a probability (number between 0 and 1). To be definite, let h be the probit link, $h(t) = \Phi(t)$, where Φ is the cumulative distribution function of the standard normal distribution. The parameters a and b are to be estimated, say by maximum likelihood. Given the parameter estimates $\hat{a}$ and $\hat{b}$, we then solve $T = \Phi(\hat{a} + \hat{b} \log C_T)$ to get

$$\hat{C}_T = \exp\big((\Phi^{-1}(T) - \hat{a})/\hat{b}\big). \tag{6.4}$$

This $\hat{C}_T$ is called a probit regression estimator.

By equation (6.2), we know that the correct value of K is $1/3$. So the first regression estimator, which we denote $\hat{C}_T^{\text{probit-(1/3)-}T}$, fits model (6.3) with $K = 1/3$ and sets $\hat{C}_T^{\text{probit-(1/3)-}T} = \hat{C}_T$ from (6.4).

Although $K = 1/3$ is the correct value in (6.3), estimating a and b in (6.3) is numerically more difficult when $K \neq 0$. A simpler alternative is to fit the model with $K = 0$, then adjust for guessing after the fact by using $P = (1/3) + (2/3)T$ in place of T in (6.4). This is the approach of Stocking et al. (2000). We denote this estimator $\hat{C}_T^{\text{probit-0-}P}$.

A third approach is simply to ignore the effect of guessing altogether: fit (6.3) with $K = 0$, and let $\hat{C}_T^{\text{probit-0-}T} = \hat{C}_T$ in (6.4).

A numerical modification is required for the probit regression estimators. When the data $(y_1, \ldots, y_8)$ are symmetric (e.g. (1, 0, 1, 1, 1, 1, 0, 1)), the estimate of the slope parameter b is either zero or very small, so that C_T in (6.4) is either very large or undefined. In these cases we set $\hat{C}_T$ equal to the ASTM threshold estimator. This is a significant modification to the single-subject regression estimators, but it is less important with some of the population threshold estimators in Section 6.2.

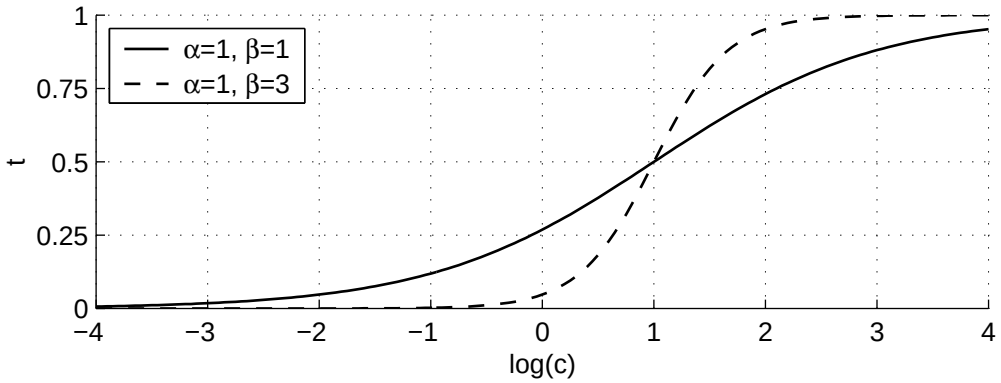

Figure 6.1: Two examples of the logistic dose-response function, as in (6.5).

6.1.2 Results

In order to evaluate the threshold estimators defined above, we consider a subject whose true, unknown odor sensitivity is defined by

$$\begin{aligned} y &\sim \text{Bernoulli}(p) \quad \text{indep.} \\ p &= 1/3 + (2/3)t \\ t &= g(\beta(\log c - \alpha)) \\ g(x) &= e^x/(1+e^x) \end{aligned} \tag{6.5}$$

for some α and β of our choosing. g is known as the logistic link function. For this subject, the 50% detection threshold is $C_{0.5} = e^{\alpha}$, and β is proportional to the slope of the dose-response curve at $C_{0.5}$. Figure 6.1 shows two examples of the logistic dose-response function. Note that while the true dose-response function is logistic, the probit threshold estimators assume a probit link, which has lighter tails. In this way we build in some misspecification of the dose-response model.

We suppose that the experimental log-concentrations are $-3.5, -2.5, \ldots, 2.5, 3.5$. This means that if α is close to zero (and if β is large enough), most of the subject's change in response occurs in the observable range, so the dose-response should be easy to estimate. If α is large or small, the dose-response changes more outside the range of observation and so should be harder to estimate. The same is true if β is small. If β is large, then the response changes within a narrow range and may be easier to observe and estimate.

Under the above assumptions, we can compute exact distributions of the various threshold estimators. At each of the 8 concentrations the subject can answer correctly or incorrectly, so there are only $2^8 = 256$ possible outcomes. For each possible outcome, we compute each of the four estimators, and also the probability of the outcome under assumptions (6.5), given α and β. The distribution of each estimator is therefore known and can be summarized in terms of means and variances, for example.

The probit estimators were computed in Matlab (The MathWorks, Inc., 1996) by the

method of maximum likelihood (McCullagh and Nelder, 1989). The likelihoods were maximized using the `fminunc` function in Matlab's Optimization Toolbox (The MathWorks, Inc., 1999).

Figures 6.2–6.9 show the results. Figures 6.2–6.5 are for estimators of $C_{0.5}$, and Figures 6.6–6.9 are for $C_{0.95}$.

Consider first the estimators of $C_{0.5}$. Figures 6.2 and 6.3 show the *log-bias*, defined as $E \log \hat{C}_{0.5} - \log C_{0.5}$, and *log-variance*, $\mathrm{Var}(\log \hat{C}_{0.5})$, of the ASTM and probit estimators of $C_{0.5}$, over a range of α and β. The performance of all four estimators depends on how well the experiment covers the subject's range of response (α) and on the steepness of the response (β). In terms of log-bias, the ASTM, probit-$(1/3)$-T, and probit-0-P estimators perform about equally well, and all have small log-bias around $\alpha = 0$, where conditions are favorable to the experiment. On the other hand the probit-0-T estimator underestimates the threshold, presumably because it mistakes guessing for detection. However, Figure 6.3 shows that any differences in log-bias are swamped by the much greater log-variance of the probit estimators. The probit-$(1/3)$-T estimator is the most variable of all, reflecting the greater difficulty of estimating a and b in (6.3) when $K \neq 0$. Thus the ASTM estimator performs best in a single-subject experiment. The probit estimators apparently suffer from having to fit a two-parameter model to only 8 binary observations.

An interesting question about the ASTM estimator is: what does it estimate? ASTM (1995b) claims to estimate a 50% threshold, but provides no evidence for the claim. Figure 6.4 answers this question, for both the ASTM and probit estimators of $C_{0.5}$. The statistic plotted in Figure 6.4 is the expected value of $\hat{T}$, where

$$\hat{T} = g(\beta(\log \hat{C}_T - \alpha)), \tag{6.6}$$

and α, β, and g are the subject's true parameters and (logistic) link function from (6.5). $\hat{T}$ is the subject's true detection probability, computed at the estimated threshold. If $\hat{T}$ is close to 0.5, then we are in fact estimating a 50% threshold. Figure 6.4 shows that on average, the ASTM estimator does indeed estimate somewhere between a 40% and 60% threshold, over a wide range of experimental conditions (α and β). Again the probit-0-T estimator tends to underestimate its target. All four estimators can grossly over- or underestimate the target if the subject's change in response occurs at the edge or outside of the range of experimental concentrations (low or high α). The slope of the change in response is less crucial.

Figure 6.5 plots the variance of $\hat{T}$ for the four estimators. The extreme variability of the probit estimators in Figure 6.3 is not apparent on the probability scale, since $\hat{T}$ is restricted to lie between 0 and 1. Even so the ASTM estimator is significantly less variable than the probit estimators when α is large.

Figures 6.6–6.9 show the same statistics for estimators of $C_{0.95}$. The ASTM estimator is the same as in the previous figures, since it does not depend on T. We saw above that $\hat{C}_T^{\mathrm{ASTM}}$ is a fairly good estimator of $C_{0.5}$, so this estimator is of course negatively biased for $C_{0.95}$. The probit estimators are less biased, but still mostly underestimate the threshold. All are about equally biased: when the detection rate is 95%, the effect of guessing is small so the probit-0-

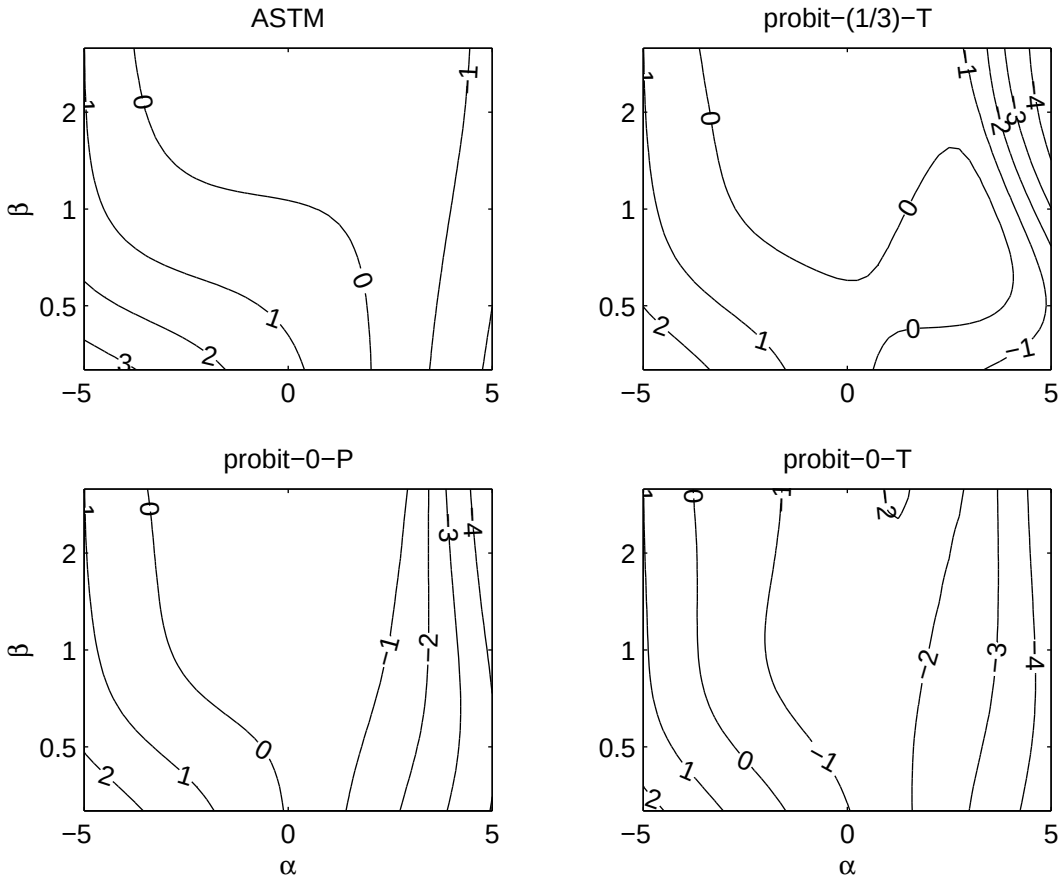

Figure 6.2: Log-bias of estimators of $C_{0.5}$

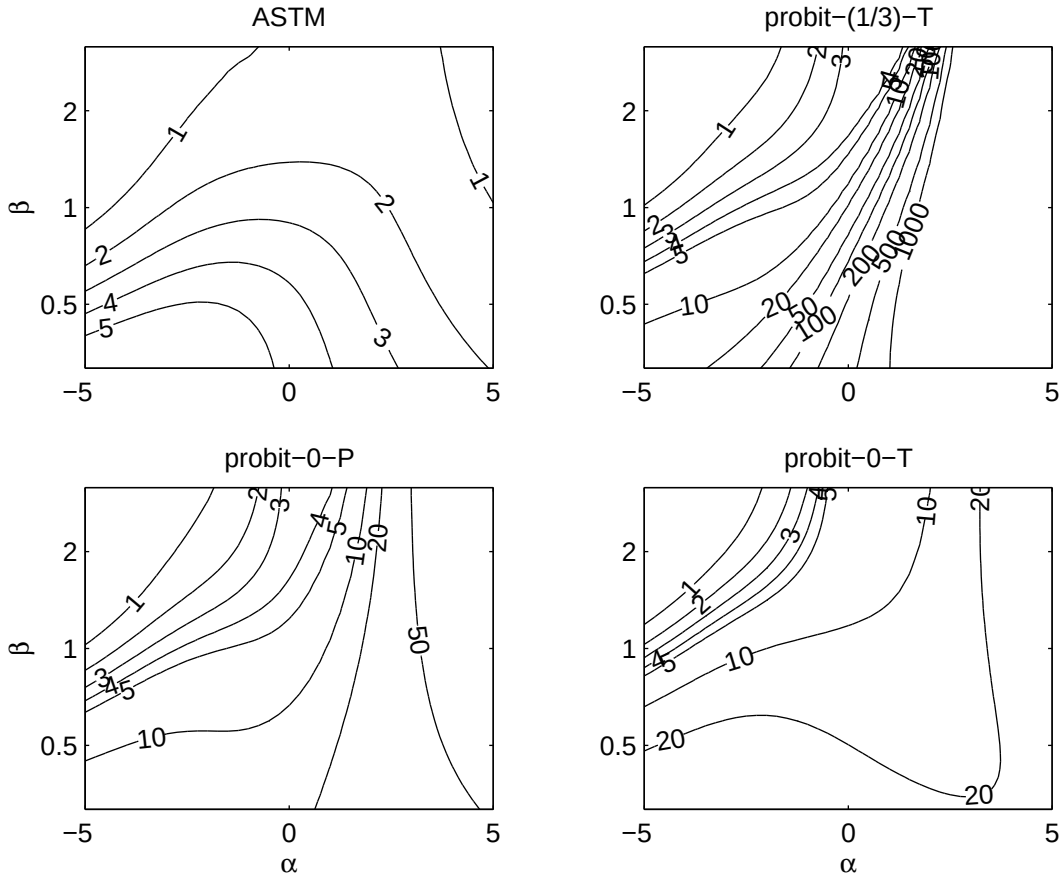

Figure 6.3: Log-variance of estimators of $C_{0.5}$

T estimator does not suffer from ignoring it. As before, $\hat{C}_T^{\text{ASTM}}$ is much less variable than any of the probit estimators, and the probit-$(1/3)$-T estimator is the most variable.

In summary, the ASTM estimator performs remarkably well as an estimator of 50% subject thresholds. It has about the same bias and much less variability than the model-based probit estimators. Of the probit estimators, the probit-0-P estimator, which neglects the effect of guessing in the model but corrects for it after the fact, is best because of its smaller bias in some cases and smaller variability than the probit-$(1/3)$-T estimator. The latter is extremely variable and should not be used in experiments of this size.

This superiority of the ASTM estimator is surprising, since it uses only part of the data and does not explicitly take guessing into account. The probit estimators might be expected to perform better in experiments with more data, for example with multiple subjects with an assumed common slope β, or with multiple replicates per subject. We consider the case of more subjects in the next Section.

For estimation of 95% subject thresholds, no estimator performs very well. The ASTM estimator estimates 50% thresholds, so it has significant negative bias. The probit estimators

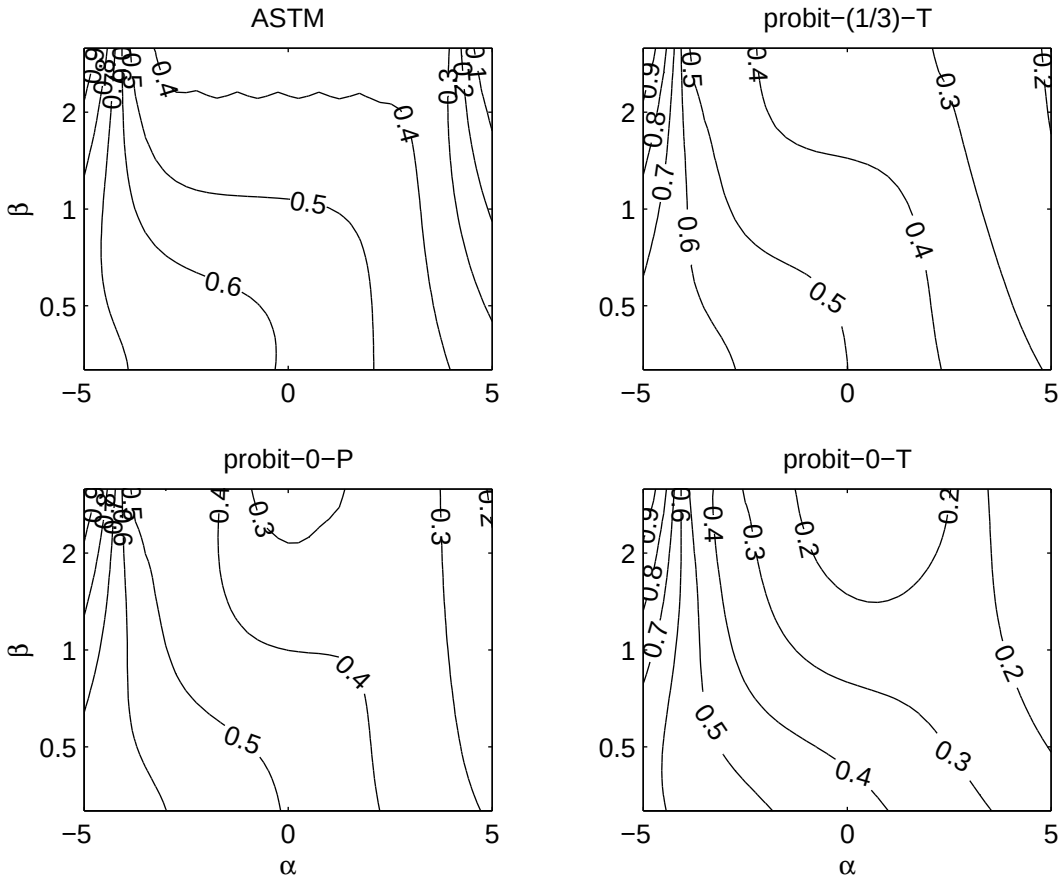

Figure 6.4: Mean true detection probability at estimates of $C_{0.5}$

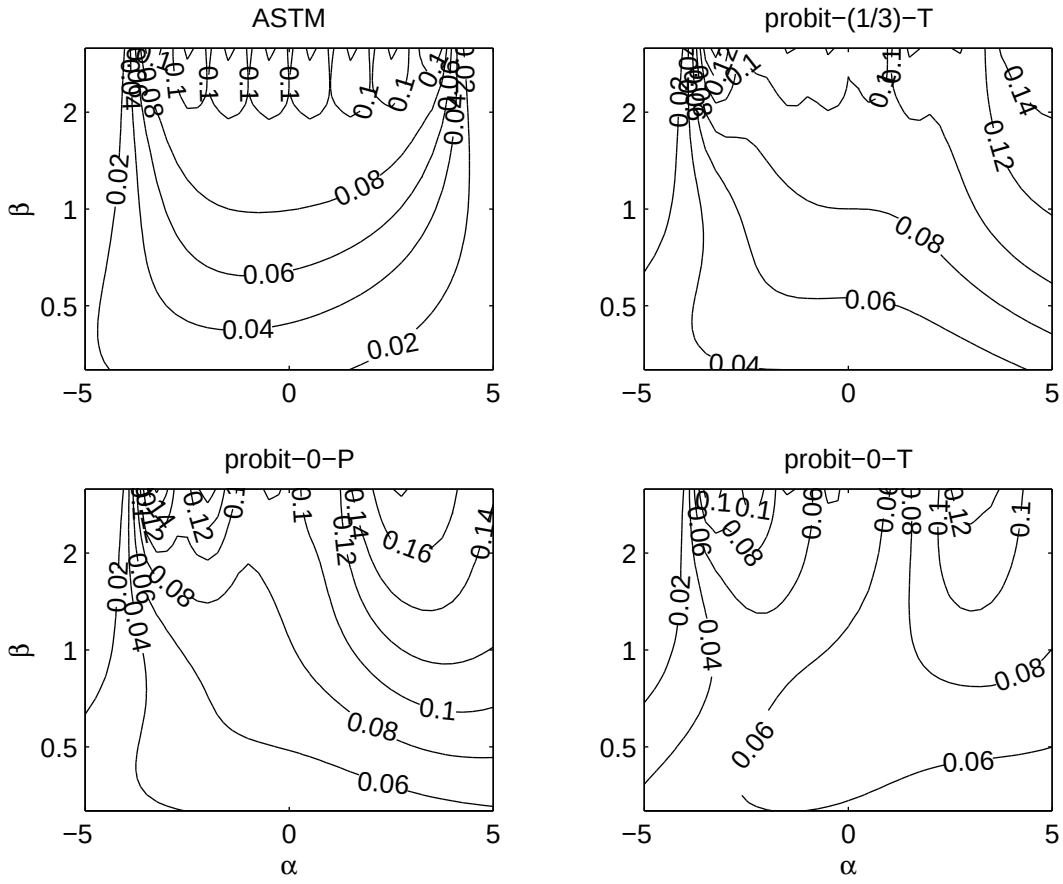

Figure 6.5: Variance of true detection probability at estimates of $C_{0.5}$

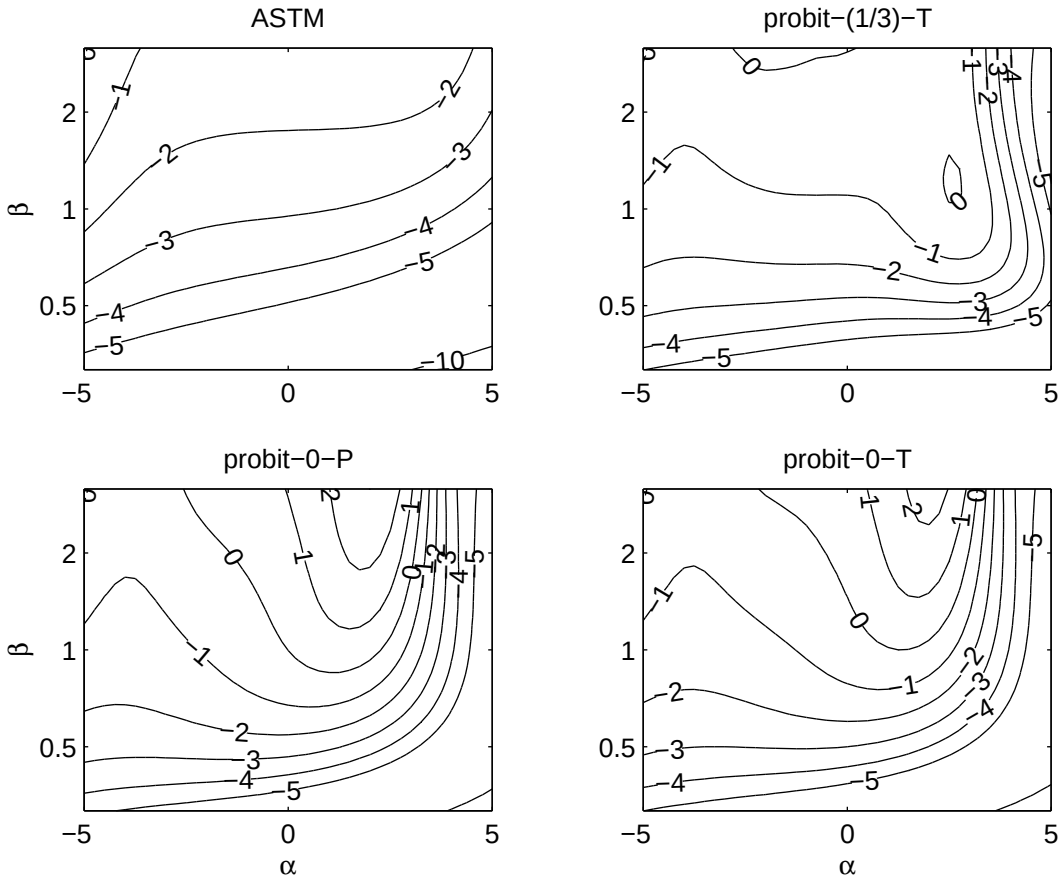

Figure 6.6: Log-bias of estimators of $C_{0.95}$

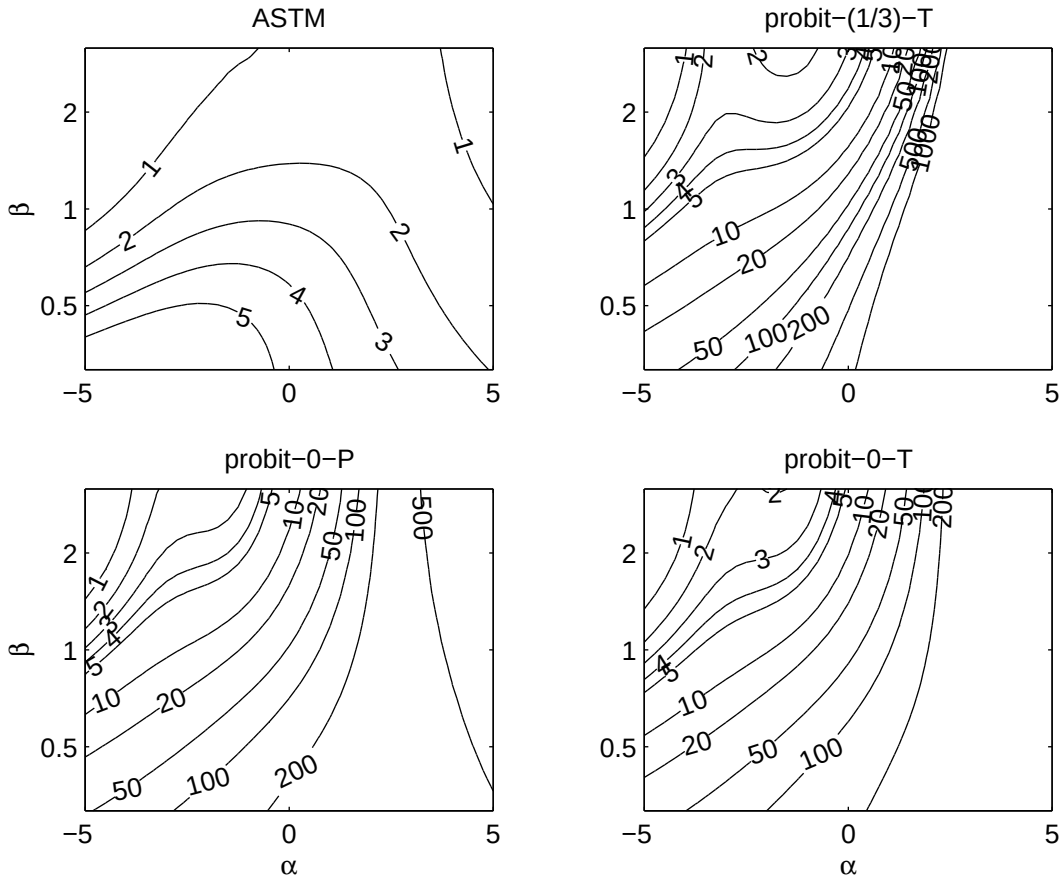

Figure 6.7: Log-variance of estimators of $C_{0.95}$

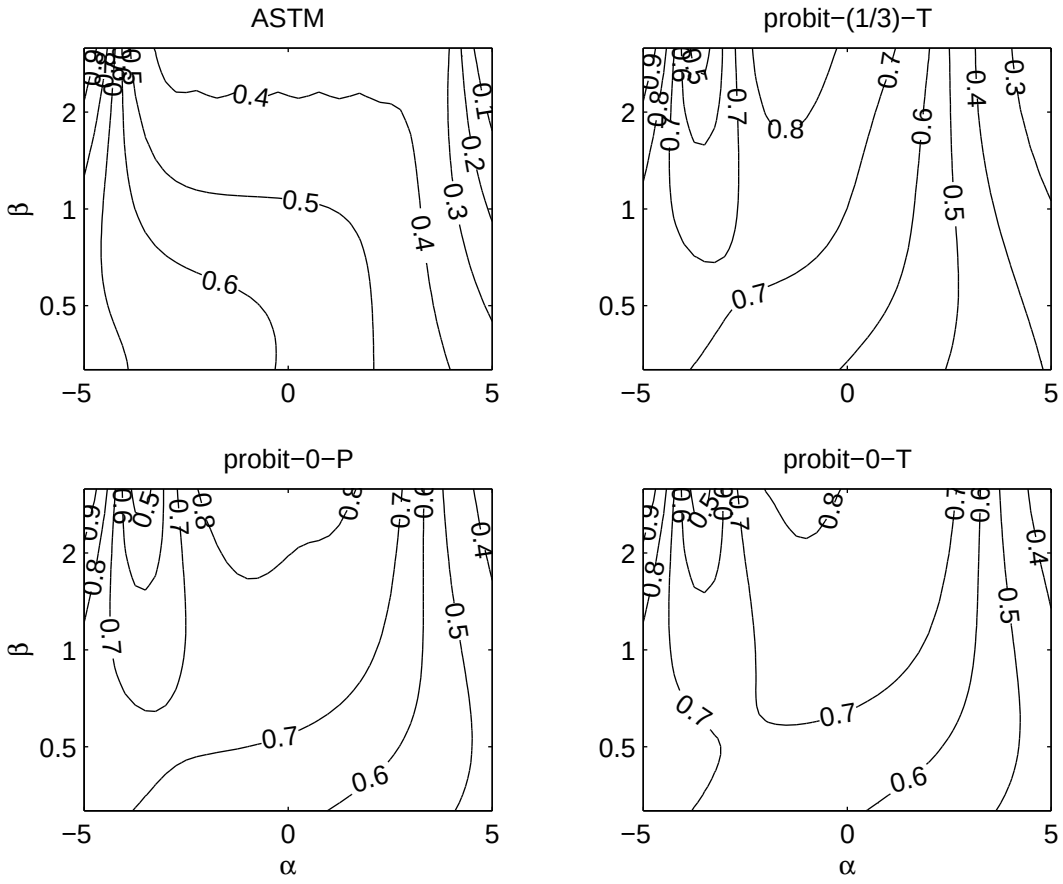

Figure 6.8: Mean true detection probability at estimates of $C_{0.95}$

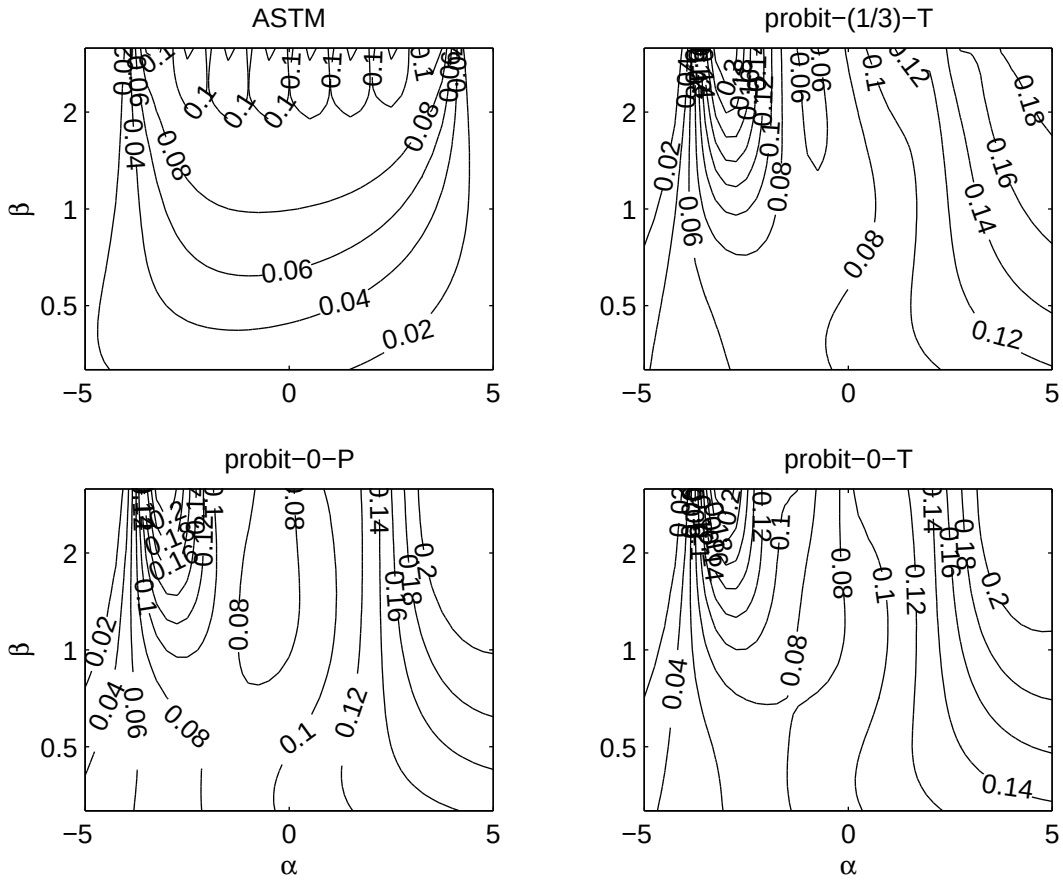

Figure 6.9: Variance of true detection probability at estimates of $C_{0.95}$

are less biased, at the cost of much higher variability. This probably just reflects the difficulty of estimating a 95% odor threshold from 8 binary responses. In experiments of this size one is probably better off restricting attention to 50% thresholds.

6.2 Population Thresholds

Consider now the same experiment as in the previous Section, but with 57 subjects. The same concentrations $c_1, \ldots, c_8$ are presented to each subject. Let

$$\begin{aligned} t_{ij} &= P(\text{subject } i \text{ detects concentration } c_j) \\ p_{ij} &= P(\text{subject } i \text{ chooses the correct bottle with concentration } c_j) \\ y_{ij} &= \begin{cases} 1 & \text{if subject } i \text{ chooses the correct bottle with concentration } c_j, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{6.7}$$

Then as before, $p_{ij} = (1/3) + (2/3)t_{ij}$ and $E(y_{ij}) = p_{ij}$.

The goal of the experiment is to estimate $C_{S,T}$, the concentration at which $(100S)\%$ of subjects can detect the contaminant at least $(100T)\%$ of the time.

6.2.1 Estimators

One way to estimate population thresholds is first to estimate a subject threshold for each subject, then compute a quantile estimate from the set of subject thresholds. We consider three estimators of this type, using different subject threshold estimators and distributional fits. A fourth estimator uses an estimate of the between-subject variability, computed in the same model as the within-subject dose-response.

For the first estimator, we compute ASTM threshold estimates $\hat{C}^{\text{ASTM}}_{T,1}, \ldots, \hat{C}^{\text{ASTM}}_{T,57}$ for the 57 subjects as in Section 6.1, then let $\hat{C}^{\text{ASTM}}_{S,T}$ be an estimate of the $(100S)$-th percentile of the subject thresholds. Quantiles can be estimated in many ways: for example as quantiles of the empirical CDF, of a kernel density estimate, or of a normal distribution fitted in any of several ways. For data that are close to normally distributed, the differences are minor. For simplicity, we use a normal quantile estimator applied to the log-ASTM thresholds:

$$\hat{C}^{\text{ASTM-normal}}_{S,T} = \exp\left(\hat{m} + \hat{s}\Phi^{-1}(S)\right), \tag{6.8}$$

where $\hat{m}$ and $\hat{s}$ are the sample mean and standard deviation, respectively, of the log-ASTM thresholds.

A problem with the ASTM thresholds is that when a subject correctly identifies the contaminant at all concentrations, or fails to identify it at the highest concentration, the threshold estimate is arbitrarily placed just outside the range of tested concentrations. In these cases the threshold might be better assumed to be censored, that is, unobserved but known to be outside the tested range. A normal distribution can then be fit to the thresholds in a way that uses the censoring information. Kroll and Stedinger (1996) compared several such methods, and found

that maximum likelihood (Cohen, 1991) works best when the goal is to estimate 10th percentiles. We therefore define a second threshold estimator, $\hat{C}_{S,T}^{\text{ASTM-censored}}$, in the same way as $\hat{C}_{S,T}^{\text{ASTM-normal}}$, but with $\hat{m}$ and $\hat{s}$ estimated by maximum likelihood, treating ASTM thresholds outside the tested range as censored. In cases where fewer than three thresholds are uncensored, the likelihood estimates are either unavailable or likely to be too variable, and so instead we let $\hat{C}_{S,T}^{\text{ASTM-censored}} = \hat{C}_{S,T}^{\text{ASTM-normal}}$.

A third estimator uses probit regression, as in Section 6.1, to estimate the subject thresholds. Since the probit estimators were found to be highly variable with only one subject, we try to benefit from multiple subjects here by estimating a common slope parameter for all subjects. The model is

$$\begin{aligned} y_{ij} &\sim \text{Bernoulli}(p_{ij}) \quad \text{indep.} \\ p_{ij} &= K + (1-K)t_{ij} \\ t_{ij} &= \Phi(a_i + b \log c_j) \end{aligned} \tag{6.9}$$

where $a_1, \ldots, a_{57}$ and b are to be estimated. Under (6.9), each subject has his or her own intercept a_i, but all subjects are assumed to share a common slope b. While this assumption is certainly not true, it allows us to borrow strength across subjects in estimating an average slope, thereby reducing the variability of the threshold estimates.

The results of Section 6.1.1 showed that the best way to account for guessing in the forced-choice test is to set $K = 0$ and estimate the i-th subject's $(100T)$% threshold as

$$\hat{C}_{T,i}^{\text{probit-0-}P} = \exp\bigl(\Phi^{-1}((1/3) + (2/3)T) - \hat{a}_i)/\hat{b}\bigr) \tag{6.10}$$

The (S, T) population threshold is then estimated as the normal-theory $(100S)$% quantile of $\hat{C}_{T,1}^{\text{probit-0-}P}, \ldots, \hat{C}_{T,57}^{\text{probit-0-}P}$, as in (6.8). We denote this estimator $\hat{C}_{S,T}^{\text{probit-fixed}}$, where the designation "fixed" stands for "fixed effects": the unknowns a_i and b in (6.9) are treated as fixed, unknown constants, in contrast to the next estimator.

A problem with model (6.9) is that it neglects the correlation that exists between repeated observations on a single subjects. Also because it treats the subject effects a_i as fixed numbers, rather than as a random sample from a population of subject effects, it does not allow inference to the population from which the subjects were selected. A model which avoids these problems is

$$\begin{aligned} (y_{ij} | p_{ij}) &\sim \text{Bernoulli}(p_{ij}) \quad \text{indep.} \\ p_{ij} &= t_{ij} + (1 - t_{ij})K \\ t_{ij} &= \Phi(a_i + b \log c_j) \\ a_i &\sim N(\mu_a, \sigma_a^2) \quad \text{indep.} \end{aligned} \tag{6.11}$$

Here the subject effects a_i are realizations of a random variable, which describes the distribution of subjects' sensitivities to odor. Allowing the subject effects to be random accomplishes two goals: it induces a correlation among the responses $y_{i1}, \ldots, y_{i8}$ on a single subject; and by taking the randomness across subjects into account, it allows inference to the population from which the subjects were drawn.

When subjects are selected by random sampling, the population (S, T) odor threshold is determined by the equation

$$P\big(\text{subject can detect concentration } c \text{ in } (100T)\% \text{ or more of samples}\big) = S \tag{6.12}$$

where the probability is with respect to a randomly selected subject. Under model (6.11), (6.12) may be solved for c to give

$$C_{S,T} = \exp\left((\Phi^{-1}(T) - \mu_a - \sigma_a \Phi^{-1}(1-S))/b\right) \tag{6.13}$$

As before to account for guessing, we choose $K = 0$ in (6.11) and substitute $P = (1/3) + (2/3)T$ for T after the fact. Given parameter estimates $\hat{\mu}_a$, $\hat{\sigma}_a^2$, and $\hat{b}$, we substitute into (6.13) to get a threshold estimate

$$\hat{C}_{S,T}^{\text{probit-mixed}} = \exp\left((\Phi^{-1}((1/3) + (2/3)T) - \hat{\mu}_a - \hat{\sigma}_a \Phi^{-1}(1-S))/\hat{b}\right) \tag{6.14}$$

The designation "mixed" in $\hat{C}_{S,T}^{\text{probit-mixed}}$ refers to model (6.11), which is said to be a "mixed model" because it contains both fixed effects (b) and random effects (a_i) to be estimated.

Estimation of the parameters in (6.11) is nontrivial. Proposed methods include pseudo- and quasi-likelihoods (Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993) and various kinds of simulated likelihoods (McCulloch, 1994; Geyer, 1994; Geyer and Thompson, 1992). We use the restricted pseudo-likelihood method of Wolfinger and O'Connell (1993), mainly because it has been implemented in a macro, `GLIMMIX`, for the SAS system (SAS Institute Inc., 1989). We used `GLIMMIX` to estimate the parameters of (6.11) in SAS.

6.2.2 Results

In order to evaluate the population threshold estimators defined above, we consider a population of 57 subjects whose true, unknown odor sensitivites are defined by

$$\begin{aligned}
(y_{ij} \mid p_{ij}) &\sim \text{Bernoulli}(p_{ij}) \quad \text{indep.} \\
p_{ij} &= 1/3 + (2/3)t_{ij} \\
t_{ij} &= g(\beta_i(\log c_j - \alpha_i)) \\
g(x) &= e^x/(1+e^x) \\
\alpha_i &\sim N(\mu_\alpha, \sigma_\alpha^2) \quad \text{indep.} \\
\beta_i &\sim LN(\mu_\beta, \sigma_\beta^2) \quad \text{indep.} \\
\alpha_i, \beta_i &\ \text{ indep.}
\end{aligned} \tag{6.15}$$

where $LN(\mu_\beta, \sigma_\beta^2)$ is the lognormal distribution with mean μ_β and variance σ_β^2 (of β, not of $\log\beta$). We must choose μ_α, μ_β, σ_α, and σ_β. Note that for subjects who follow (6.15), the probit estimators are derived from models that are misspecified in two ways: they assume a

Table 6.1: Parameter values for the population threshold simulation.

μ_α	-3	0	3
σ_α^2	0.5	2	8
μ_β	0.3	1	3
σ_β^2	0	0.25	1

probit link and a common slope β for all subjects. The results of the probit estimators will therefore include some model error.

In the single-subject experiment in the previous section, we were able to obtain exact distributions of the threshold estimators because the number of possible outcomes was small. With 57 subjects that is no longer true, so we used simulation instead to evaluate the estimators.

For the simulation we chose values of μ_α, μ_β, σ_α^2, and σ_β^2 shown in Table 6.1. These values were chosen to represent a range of difficulty for the estimation. By comparison, the estimated parameters from the Stocking et al. (2000) data, transformed to model (6.15) and scaled to the simulated predictor range, are $\mu_\alpha = 1.27$, $\mu_\beta = 0.79$, $\sigma_\alpha^2 = 2.69$, and $\sigma_\beta^2 = 0.06$. All of these estimates are well within the range of values in Table 6.1.

We used each combination of values of the four parameters, for a total of 81 parameter combinations. At each parameter combination we simulated 20 i.i.d. populations of 57 subjects following (6.15). For the experimental log-concentrations, as before we chose $-3.5, -2.5, \ldots, 2.5, 3.5$ for each subject. For each sample population we computed the ASTM-normal, ASTM-censored, probit-fixed, and probit-mixed estimators of the (0.10, 0.50), (0.10, 0.95), and (0.50, 0.50) population thresholds.

Figures 6.10—6.12 show the results of the simulation, in terms of squared log-bias, log-variance, and log-mean squared error (MSE) of estimates of the (50%, 50%), (10%, 50%), and (10%, 95%) thresholds. The simulation yields a large number of results, with 4 estimators of each of the 3 estimands at each of 81 parameter combinations. In order to make the results easier to digest, we summarized them first by averaging over the tested values of σ_α and σ_β. Both σ_α and σ_β turned out to have only a small effect on bias and variance, at least within the range of values that we tested.

Figures 6.10—6.12 lead to three conclusions. First, thresholds with S and T close to $1/2$ are easiest to estimate. In the case of T this agrees with the results of the previous section. Even with 57 subjects, it is difficult to estimate 95% subject thresholds from only 8 binary observations on each subject.

Second, no matter which threshold is being estimated, the results depend heavily on the experimental conditions. They depend especially on μ_β, the mean slope of the dose-response curves. All four estimators perform best when μ_β is large, and badly when μ_β is close to zero. Geometrically, when μ_β is small the dose-response curve is relatively flat, so the inverse calculation used to compute the threshold is sensitive to misspecification of the curve. Algebraically, in (6.14) this happens because $\hat{b}$ is small in the denominator.

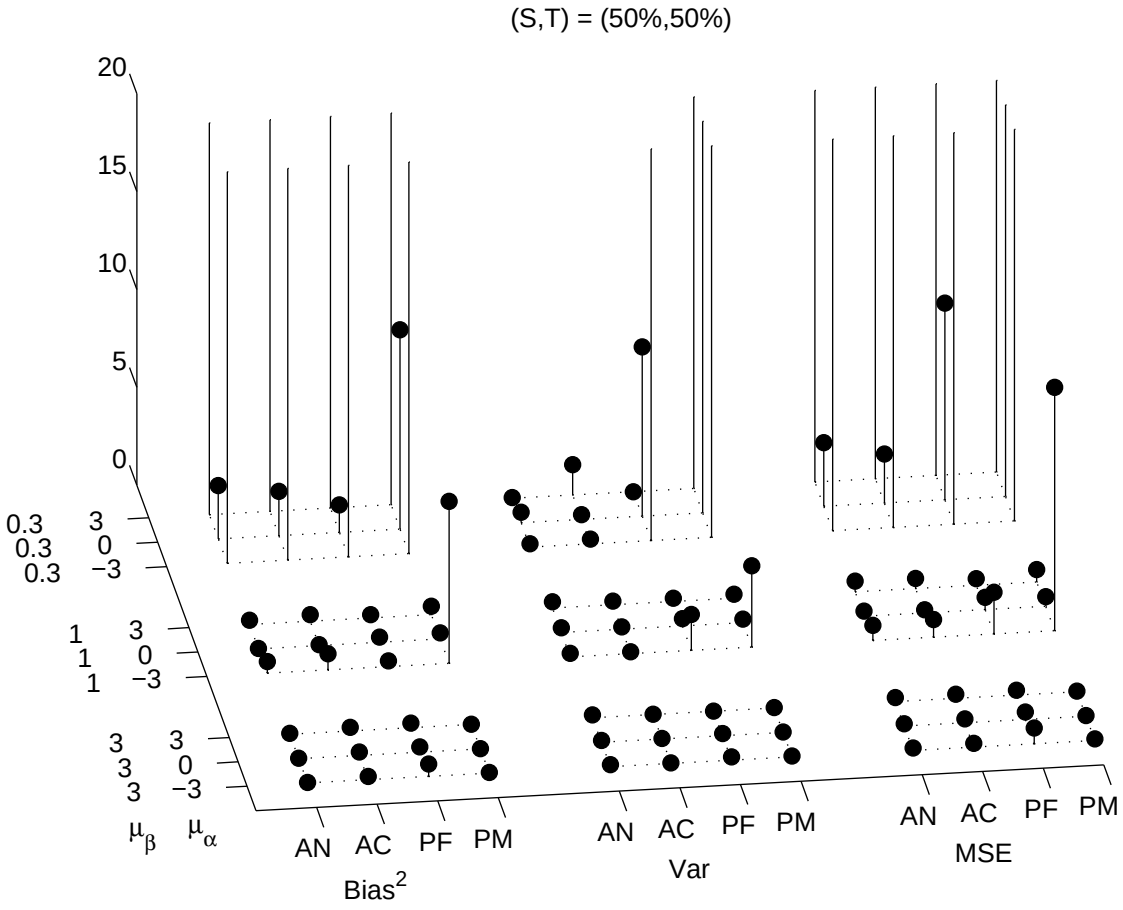

Figure 6.10: Squared log-bias, log-variance, and log-MSE of the ASTM-normal (AN), ASTM-censored (AC), probit-fixed (PF), and probit-mixed (PM) estimators of $C_{0.5,0.5}$. Responses are averaged over the simulated values of σ_α and σ_β. Where circles are missing, the response is off the scale of the graph.

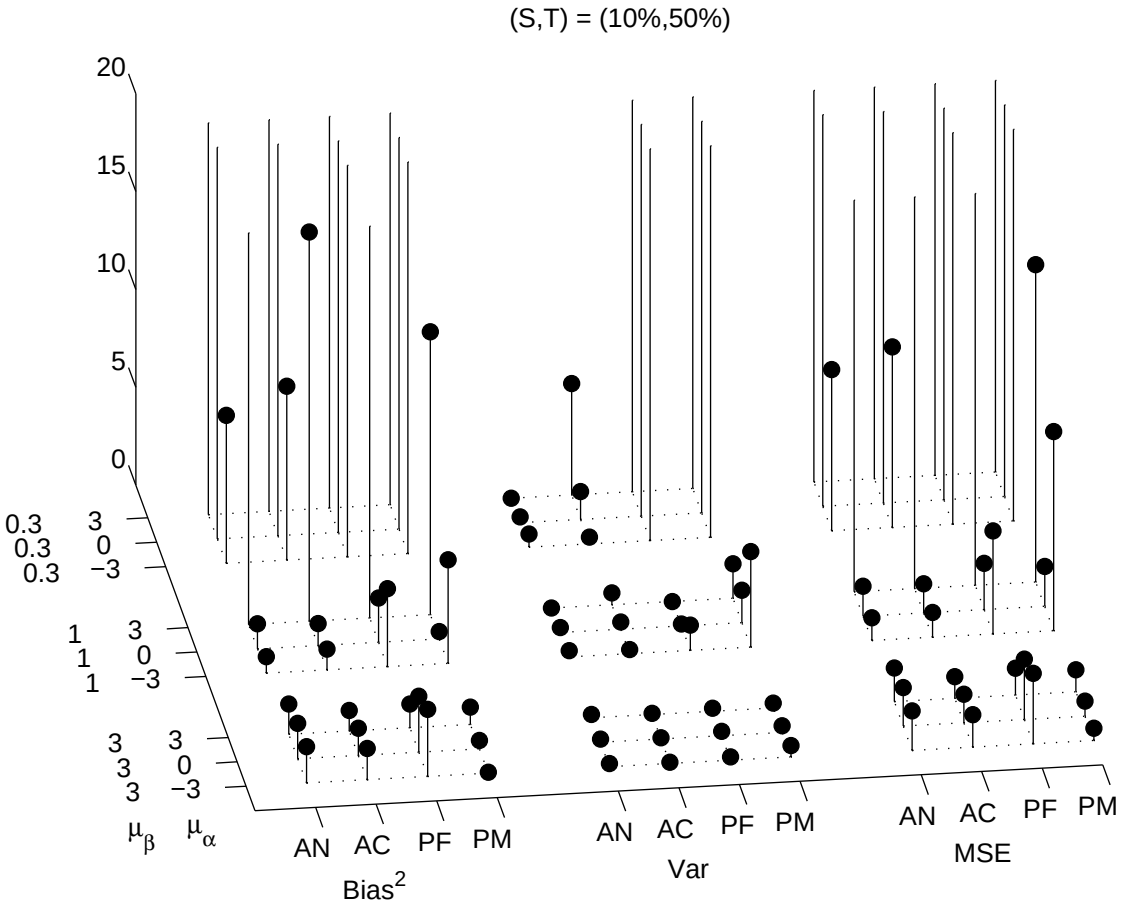

Figure 6.11: Squared log-bias, log-variance, and log-MSE of the ASTM-normal (AN), ASTM-censored (AC), probit-fixed (PF), and probit-mixed (PM) estimators of $C_{0.1,0.5}$. Responses are averaged over the simulated values of σ_α and σ_β. Where circles are missing, the response is off the scale of the graph.

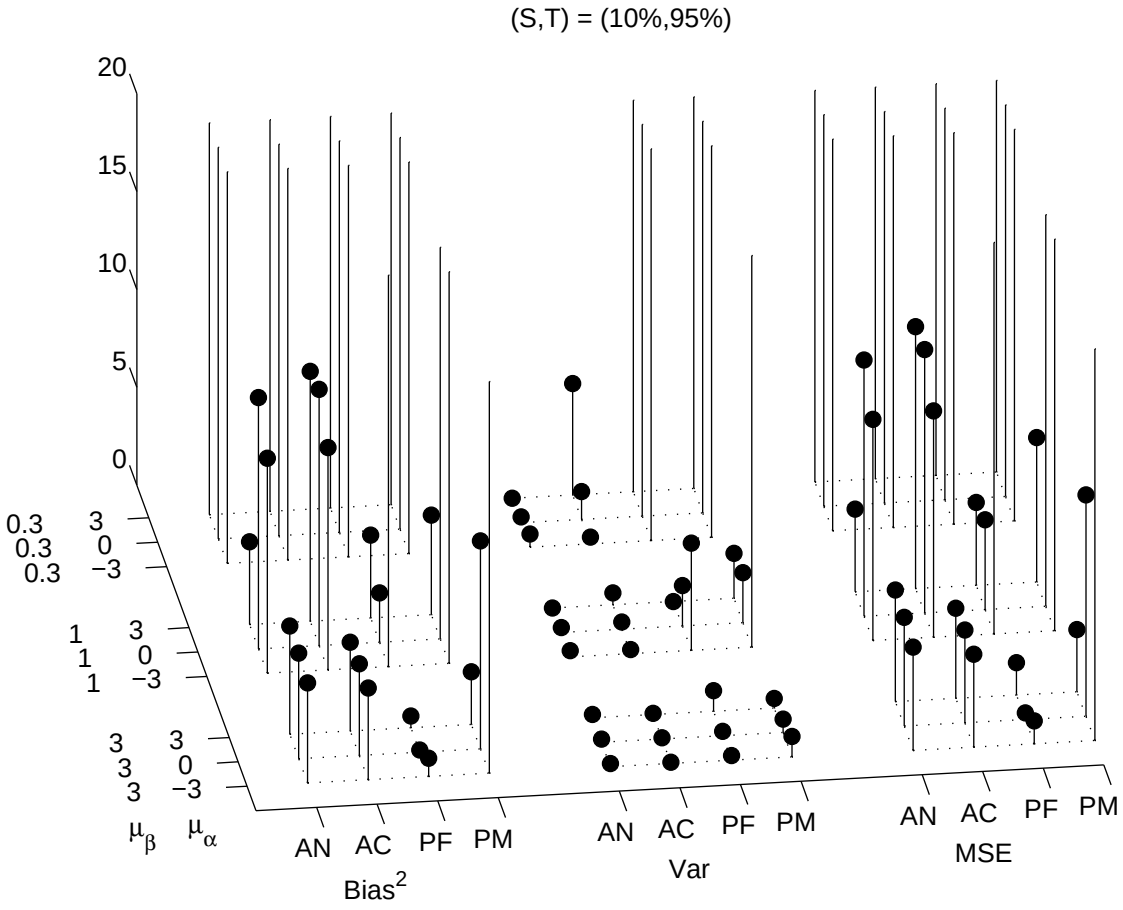

Figure 6.12: Squared log-bias, log-variance, and log-MSE of the ASTM-normal (AN), ASTM-censored (AC), probit-fixed (PF), and probit-mixed (PM) estimators of $C_{0.1,0.95}$. Responses are averaged over the simulated values of σ_α and σ_β. Where circles are missing, the response is off the scale of the graph.

Third, the ASTM-normal estimator has small variance in all cases, and this makes it a good choice for thresholds with T close to 0.5, where it also has small bias. When $T = 0.95$, no estimator performs uniformly best; the probit-fixed estimator is usually best but can also perform quite badly, for example when $\mu_\beta = 1$ and $\mu_\alpha = 3$.

The ASTM-censored estimator performs generally about the same as the ASTM-normal. Its bias is smaller in some cases and larger in others; variance is generally low but rises in some difficult cases. Although perhaps theoretically more satisfactory, the ASTM-censored estimator yields no clear advantage to compensate for its extra computational complexity.

The probit-mixed estimator performs the worst of the four, with generally higher bias and variance than the ASTM and probit-fixed estimators. However, this estimator may still be useful because it can provide confidence intervals that take into account both inter- and intra-subject variability and intra-subject correlation. By contrast, the ASTM estimators come with no subject-specific confidence intervals at all. One can form a confidence interval for the population quantile estimate, but this approach neglects intra-subject variability. For the probit-fixed estimator, the probit regression provides confidence intervals for subject thresholds, which must then be combined across subjects, taking into account the normal quantile estimate. This approach also neglects intra-subject correlation. However, we have not evaluated the accuracy of confidence intervals from any of the threshold estimators.

In the case of Stocking et al. (2000), the estimated parameters, scaled to the simulated predictor range, are $\hat{\mu}_\alpha = 1.27$ and $\hat{\mu}_\beta = 0.79$. Figure 6.10 shows that each of the three estimators may perform either well or badly when μ_β is between 0.3 and 1 and μ_α is between 0 and 3. The ASTM-normal estimator is probably best in this case, since it at least has small variability throughout that range.

7. Conclusion

Our analysis leads to the following conclusions.

1. Odor detection thresholds should be defined as the concentrations at which a certain percent of people can detect the contaminant a certain percent of the time. Both the time and subject fractions must be specified in order for a threshold to be interpretable.

2. In experiments with small amounts of data on each subject (e.g. 8 presentations and no replicates), the simple threshold estimator specified in ASTM method E679-91 performs at least as well as, and often better than, more complicated estimators based on probability regression. With this small amount of data, thresholds with a time fraction of 50% are easiest to estimate.

3. On average, the ASTM E679-91 threshold estimator estimates a subject's 40%–60% detection threshold, over a wide range of experimental conditions.

4. Each of the previous odor threshold studies of MTBE has one or more methodological problems. The data from Stocking et al. (2000), however, are sound and represent the largest and best-designed study of MTBE odor thresholds to date.

5. For design of an odor detection experiment in a way that facilitates statistical analysis, ASTM method E679-91 is preferable to Standard Method 2150B. The ASTM method, because it uses a forced-choice test, allows guessing to be treated statistically in a straightforward manner. Under the Standard Method, statistical analysis is difficult because the probabilities of guesses are unknown and may depend on the subject's knowledge of the experimental protocol.

6. Based on a reanalysis of the data from Stocking et al. (2000), 50% of people can detect MTBE at least 50% of the time in drinking water at a concentration of 15 ppb, or between 10 and 22 ppb with 95% confidence. At 5 ppb, about 23% of subjects can detect MTBE at least half the time in drinking water, or between 14% and 32% of subjects with 95% confidence.

Significant uncertainties remain in our threshold estimates. The estimates may be too low, because smokers and older subjects were excluded from Stocking et al.'s experiment. Also the relationship between detecting odor on the one hand, and rejecting a sample of water as undrinkable on the other, is complex and outside the scope of this study.

Appendix A. Individual Response Data from Stocking et al. (2000)

The following table of individual subject responses is reproduced from Stocking et al. (2000), Table 7. Correct and incorrect responses are indicated by + and o, respectively. The rightmost column lists the threshold estimates computed according to ASTM method E679-91, described in Section 3.1. The bottom row shows the total number of correct responses.

	MTBE Concentration (ppb)								ASTM
Subject	2	3.5	6	11	19	33	57	100	threshold
1	+	o	+	+	+	+	+	+	4.6
2	o	o	o	+	+	+	+	+	8.1
3	o	+	o	o	o	o	+	+	43.2
4	o	o	o	o	o	o	+	+	43.2
5	+	o	+	o	+	+	o	+	75.6
6	o	o	+	+	+	+	+	+	4.6
7	o	+	+	o	+	+	+	+	14.1
8	+	+	o	+	+	+	+	+	8.1
9	o	o	+	o	+	+	+	+	14.1
10	o	o	o	o	o	o	o	+	75.6
11	+	o	+	+	+	+	+	+	4.6
12	o	o	o	o	+	o	+	o	132.3
13	+	+	+	+	+	+	+	+	1.5
14	+	+	+	+	+	+	+	+	1.5
15	o	+	+	o	+	+	+	+	14.1
16	o	o	+	o	o	o	+	o	132.3
17	+	o	+	+	+	+	+	+	4.6
18	o	o	+	+	+	+	+	+	4.6
19	+	+	+	+	+	+	+	+	1.5
20	+	o	+	+	o	+	o	o	132.3
21	+	+	+	+	+	+	+	+	1.5
22	+	+	+	+	+	+	+	+	1.5
23	+	+	o	o	+	+	+	+	14.1
24	+	+	+	+	+	+	+	+	1.5

continued on next page

	MTBE Concentration (ppb)								ASTM
Subject	2	3.5	6	11	19	33	57	100	threshold
	continued from previous page								
25	o	+	+	+	o	+	+	+	24.7
26	o	+	+	+	o	+	+	+	24.7
27	+	o	o	o	+	o	o	+	75.6
28	o	o	+	+	+	+	+	o	132.3
29	o	+	o	o	+	+	+	+	14.1
30	o	+	+	o	+	+	+	+	14.1
31	+	o	o	+	o	o	+	o	132.3
32	+	+	+	+	o	o	o	+	75.6
33	+	+	+	o	+	o	+	+	43.2
34	o	o	o	o	o	o	+	o	132.3
35	o	o	o	+	o	+	+	o	132.3
36	o	o	o	+	o	+	+	+	24.7
37	+	+	+	+	+	+	+	+	1.5
38	+	o	+	+	+	+	+	+	4.6
39	+	+	o	o	+	+	+	+	14.1
40	o	o	+	o	+	+	+	+	14.1
41	+	+	+	+	+	+	+	+	1.5
42	o	+	o	+	+	+	+	+	8.1
43	o	o	o	o	o	o	+	+	43.2
44	o	+	+	o	+	o	+	+	43.2
45	o	+	+	o	+	+	+	+	14.1
46	+	+	+	+	+	o	+	+	43.2
47	o	+	o	o	o	+	+	o	132.3
48	o	+	o	+	o	+	+	+	24.7
49	o	o	o	+	o	+	+	+	24.7
50	o	o	+	+	+	+	+	+	4.6
51	o	o	o	+	+	+	+	+	8.1
52	+	+	+	+	+	+	+	+	1.5
53	+	+	+	+	+	+	+	+	1.5
54	+	o	o	o	o	+	+	+	24.7
55	o	o	+	+	+	+	+	+	4.6
56	o	o	o	o	o	+	+	+	24.7
57	o	o	+	+	o	+	o	+	75.6
Total correct	25	28	35	34	38	44	51	49	

References

APHA (1995), "Threshold odor test," Method 2150B, in *Standard Methods for the Examination of Water and Wastewater* (19th ed.), Washington, D.C.: American Public Health Association.

API (1994), "Odor threshold studies performed with gasoline and gasoline combined with MTBE, ETBE and TAME," API Publication 4592, American Petroleum Institute, Washington, D.C.

ASTM (1995a), "Standard practice for defining and calculating individual and group sensory thresholds from forced-choice data sets of intermediate size," Method E1432-91, in *1995 Annual Book of ASTM Standards*, volume 15.07, West Conshohocken, Pa.: American Society of Testing and Materials.

ASTM (1995b), "Standard practice for determination of odor and taste thresholds by a forced-choice ascending concentration series method of limits," Method E679-91, in *1995 Annual Book of ASTM Standards*, volume 15.07, West Conshohocken, Pa.: American Society of Testing and Materials.

Breslow, N. E. and Clayton, D. G. (1993), "Approximate inference in generalized linear mixed models," *Journal of the American Statistical Association*, 88, 9–25.

Cohen, A. C. (1991), *Truncated and Censored Samples: Theory and Applications*, New York: Marcel Dekker.

Dale, M. S., Moylan, M. S., Koch, B., and Davis, M. K. (1998), "MTBE: Taste-and-odor threshold determinations using the flavor profile method," in *Proceedings 1997 Water Quality Technology Conference, Denver, Colorado, November 9–13, 1997*, American Water Works Association.

Efron, B. and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, London: Chapman & Hall.

Geyer, C. J. (1994), "Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo," Technical Report 568, School of Statistics, University of Minnesota.

Geyer, C. J. and Thompson, E. A. (1992), "Constrained Monte Carlo maximum likelihood for dependent data," *Journal of the Royal Statistical Society* (ser. B), 54, 657–699.

HMSO (1982), "Methods for the examination of water and associated materials, odour and taste in raw and potable waters," Standing Committee of Analysts, HMSO, London.

Kroll, C. N. and Stedinger, J. R. (1996), "Estimation of moments and quantiles using censored data," *Water Resources Research*, 32, 1005–1012.

Malcolm Pirnie (1998), "Technical memorandum: Taste and odor properties of methyl tertiary-butyl ether and implications for setting a secondary maximum contaminant level," unpublished.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

McCulloch, C. E. (1994), "Maximum likelihood estimation of variance components for binary data," *Journal of the American Statistical Association*, 89, 330–335.

Prah, J. S., Goldstein, F. M., Devlin, R., Otto, D., Ashley, D., House, S., Cohen, K. L., and Gerrity, T. (1994), "Sensory, symptomatic, inflammatory, and ocular responses to and the metabolism of methyl tertiary-butyl ether in a controlled human exposure experiment," *Inhalation Toxicology*, 6, 521–538.

SAS Institute Inc. (1989), *SAS/STAT User's Guide, Version 6* (4th ed.), Cary, NC: SAS Institute Inc.

Schiffman, S. (1992), "Olfaction in aging and medical disorders," in *Science of Olfaction*, eds. M. J. Serby and K. L. Chobor, 500–525, New York: Springer-Verlag.

Shen, Y. F., Bergen, M., Yoo, L. J., and Fitzsimmons, S. R. (1998), "Effect of residual chlorine on the threshold odor concentrations of MTBE in drinking water," in *Proceedings 1997 Water Quality Technology Conference, Denver, Colorado, November 9–13, 1997*, American Water Works Association.

Shen, Y. F., Yoo, L. J., Fitzsimmons, S. R., and Yamamoto, M. K. (1997), "Threshold odor concentrations of MTBE and other fuel oxygenates," in *Proceedings of the 1996 American Chemical Society Meeting, San Francisco*, American Chemical Society.

Smith, D. V. and Duncan, H. J. (1992), "Primary olfactory disorders: Anosmia, hyposmia, and dysosmia," in *Science of Olfaction*, eds. M. J. Serby and K. L. Chobor, 439–466, New York: Springer-Verlag.

Stocking, A. J., Suffet, I. H., McGuire, M. J., and Kavanaugh, M. C. (2000), "Implications of a MTBE consumer threshold odor study for drinking water standard setting," Malcolm Pirnie, Inc., Oakland, Calif., draft of June 12, 2000.

The MathWorks, Inc. (1996), *Using MATLAB*, Natick, Mass.: The MathWorks, Inc.

The MathWorks, Inc. (1999), *Optimization Toolbox User's Guide*, Natick, Mass.: The MathWorks, Inc.

TRC (1993), "Final report to ARCO Chemical Company on the odor and taste threshold studies performed with methyl-tertiary-butyl ether (MTBE) and ethyl-tertiary-butyl ether (ETBE)," TRC Project No. 13442-M31, TRC Environmental Corporation, Windsor, Conn.

US EPA (2000), "MTBE in fuels," http://www.epa.gov/mtbe/gas.htm.

Wolfinger, R. and O'Connell, M. (1993), "Generalized linear mixed models: A pseudo-likelihood approach," *Journal of Statistical Computation and Simulation*, 48, 233–243.

Young, W. F., Horth, H., Crane, R., Ogden, T., and Arnott, M. (1996), "Taste and odour threshold concentrations of potential potable water contaminants," *Water Research*, 30, 331–340.